Statistics and Econometric Models

VOLUME 2

Themes in Modern Econometrics

Managing editor
PETER C.B. PHILLIPS, *Yale University*

Series editors
ADRIAN PAGAN, *Australian National University*
CHRISTIAN GOURIEROUX, *CREST and CEPREMAP, Paris*
MICHAEL WICKENS, *University of York*

Themes in Modern Econometrics is designed to service the large and growing need for explicit teaching tools in econometrics. It will provide an organised sequence of textbooks in econometrics aimed squarely at the student population, and will be the first series in the discipline to have this as its express aim. Written at a level accessible to students with an introductory course in econometrics behind them, each book will address topics or themes that students and researchers encounter daily. While each book will be designed to stand alone as an authoritative survey in its own right, the distinct emphasis throughout will be on pedagogic excellence.

Titles in the series

Statistics and Econometric Models: Volume 1
CHRISTIAN GOURIEROUX and ALAIN MONFORT
Translated by QUANG VUONG

Statistics and Econometric Models: Volume 2
CHRISTIAN GOURIEROUX and ALAIN MONFORT
Translated by QUANG VUONG

Time Series and Dynamic Models
CHRISTIAN GOURIEROUX and ALAIN MONFORT
Translated and edited by GIAMPIERO GALLO

This two-volume work aims to present, as completely as possible, the methods of statistical inference with special reference to their economic applications. It is a well-integrated textbook presenting a wide diversity of models in an entirely coherent and unified framework. The reader will find a description not only of the classical concepts and results of mathematical statistics, but also of concepts and methods recently developed for the specific needs of econometrics. The authors have sought to avoid an overly technical presentation and go to some lengths to encourage an intuitive understanding of the results by providing numerous examples throughout. The breadth of approaches and the extensive coverage of these volumes provide a thorough and entirely self-contained course in modern econometrics.

Volume 1 provides an introduction to general concepts and methods in statistics and econometrics, and goes on to cover estimation, prediction and algorithms. Volume 2 focuses on testing, confidence regions, model selection, and asymptotic theory.

Statistics and Econometric Models

VOLUME 2

Testing, Confidence Regions, Model Selection, and Asymptotic Theory

---

Christian Gourieroux
CREST and CEPREMAP, Paris
Alain Monfort
CREST–INSEE, Paris
Translated by Quang Vuong

CAMBRIDGE
UNIVERSITY PRESS

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

Originally published in French as
Statistique et modèles économétriques Vol. 2 by Économica 1989

and © Ed. ÉCONOMICA, 1989

First published in English by Cambridge University Press 1995
as *Statistics and econometric models Volume 2*

English translation © Cambridge University Press 1995

Printed in Great Britain at the University Press, Cambridge

A catalogue record of this book is available from the British Library

Library of Congress cataloguing in publication data

Gourieroux, Christian, 1949–
[Statistique et modéles économétriques. English]
Statistics and Econometric Models: Testing, Confidence Regions,
Model Selection, and Asymptotic Theory / Christian Gourieroux, Alain Monfort;
translated by Quang Vuong.
p. cm. – (Themes in Modern Econometrics 2)
Includes bibliographical references.
ISBN 0 521 47162 1 (v. II). – ISBN 0 521 47745 X (v. II : pbk.)
1. Statistics. 2. Econometric models. I. Monfort, Alain, 1943–. II. Title.
HB137.G6613 1995
330'.01'5191–dc20 94-25194
CIP

ISBN 0 521 40551 3 (V. 1) 0 521 47162 1 (V. 2)
ISBN 0 521 47744 1 (V. 1 pb) 0 521 47745 X (V. 2 pb)
ISBN 0 521 47837 5 (two volume paperback set)

ISBN 0 521 47162 1 (hardback)
ISBN 0 521 47745 X (paperback)

TAG

Contents

14 Introduction to Tests of Hypotheses *page* **1**
14.1 Testing Theory and Modelling 1
14.2 Hypotheses 2
14.3 Examples 3
14.4 Definition of a Test 6
14.5 Types of Errors 7
14.6 Ordering Hypothesis Tests 9
14.7 Choice of a Test 12
14.7.1 Bayesian Principle 13
14.7.2 Neyman Principle 14
14.8 Exercises 15
14.9 References 16

15 Uniformly Most Powerful Tests **17**
15.1 Tests of Simple Hypotheses 17
15.1.1 Risk Diagram 17
15.1.2 Neyman–Pearson Theorem 22
15.1.3 Application to Exponential Families 26
15.2 Tests of One-Sided Hypotheses 28
15.2.1 Monotone Likelihood Ratio Families 28
15.2.2 Application to Exponential Families 30
15.2.3 Locally Uniformly Most Powerful Tests and Score Tests 31
15.3 Tests of Two-Sided Hypotheses 35
15.3.1 Case Where $\Theta_0 = \{\theta : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2\}$ 36
15.3.2 Case Where $\Theta_0 = \{\theta : \theta_1 \leq \theta \leq \theta_2\}$ 36
15.4 Exercises 37
15.5 References 39

16 Unbiased Tests and Invariant Tests 41
16.1 Unbiased Tests 41
16.1.1 Unbiased Tests and α-Similar Tests 41
16.1.2 UMPU Tests in the Absence of Nuisance Parameters 43
16.1.3 UMPU Tests in the Presence of Nuisance Parameters 47
16.1.4 Locally Uniformly Most Powerful Unbiased Tests 65
16.2 Invariant Tests 70
16.2.1 Models and Problems With Invariant Tests 70
16.2.2 Invariant and Maximal Invariant Tests 71
16.2.3 Application: Fisher Tests 72
16.3 Exercises 77
16.4 References 79

17 Likelihood Based Tests 81
17.1 Wald, Score, and Likelihood Ratio Tests: Principles 81
17.2 Classical Tests of $g(\theta) = 0$ 82
17.2.1 Wald Test 84
17.2.2 The Score or Lagrange Multiplier Test 86
17.2.3 The Likelihood Ratio Test 90
17.2.4 Equivalent Tests and Independence Property 93
17.3 Examples of Tests of $g(\theta) = 0$ 96
17.3.1 The Linear Model With Known Variance 96
17.3.2 The Linear Model With Unknown Variance 98
17.3.3 The Nonlinear Gaussian Regression Model 101
17.4 Classical Tests of $\theta = h(\gamma)$ 103
17.4.1 A Wald-Type Test 103
17.4.2 Score Test 106
17.4.3 Likelihood Ratio Test 107
17.4.4 Equivalent Tests and Independence Property 108
17.5 Examples of Tests of a Null Hypothesis in Explicit Form: The Chi-Square Tests 109
17.5.1 Tests of Goodness-of-Fit to a Family of Distributions 109
17.5.2 Tests of Independence 111
17.5.3 Goodness-of-Fit Tests to a Distribution 113
17.6 Score Tests and Augmented Regressions 114
17.6.1 Omitted-Variable Tests in Nonlinear Regression Models 114
17.6.2 Augmented Regression Principle 115

17.6.3 A Homoscedasticity Test 116
17.6.4 A Test of Independence of the Errors 117
17.7 Score Tests and Generalized Residuals 119
17.7.1 Latent Exponential Model and Score Vector 119
17.7.2 Application to the Probit Model 121
17.8 Neyman or $C(\alpha)$ Tests 122
17.8.1 Principle 122
17.8.2 Example 125
17.9 Hausman (Likelihood Based) Tests 125
17.9.1 Test Description 125
17.9.2 Examples 128
17.10 Exercises 130
17.11 References 133

18 General Asymptotic Tests **135**
18.1 General Asymptotic Tests 135
18.1.1 Tests Using the Constrained Maximization of $L_n(\theta)$ 136
18.1.2 Tests Based on the Unconstrained Estimator $\hat{\theta}_n$ of θ 145
18.1.3 Tests Based on the Function $g(\hat{\theta}_n, a)$ 147
18.1.4 An Example: Tests Based on Least Squares 150
18.2 ALS Based Specification Tests 153
18.2.1 General Principle 153
18.2.2 Unknown Linear Constraints on Functions of the Parameter θ 155
18.2.3 Examples 156
18.3 GMM Based Specification Tests 160
18.3.1 General Result 160
18.3.2 Examples 163
18.4 Hausman Specification Tests 166
18.4.1 Principle 166
18.4.2 Example: An Exogeneity Test Based on an Augmented Regression 167
18.5 Information Matrix Test 170
18.6 Exercises 174
18.7 References 176

19 Multiple Tests **179**
19.1 The Linear Model 179
19.1.1 Nested Hypotheses 179
19.1.2 The Testing Procedure 180

19.1.3 Interpretation in Terms of Correlation Coefficients 183
19.2 Implicit Form: The General Case 184
19.2.1 Test of a Hypothesis When the Maintained Hypothesis is Already Constrained 184
19.2.2 Sequential Tests of Nested Hypotheses in Implicit Form 195
19.2.3 Example 197
19.3 Explicit Form: The General Case 198
19.3.1 Test of a Hypothesis When the Maintained Hypothesis is Already Constrained 198
19.3.2 Sequential Tests of Nested Hypotheses in Explicit Form 200
19.3.3 Example: Tests for a Common Root 200
19.4 General Multiple Tests 201
19.4.1 Tests of Nonnested Null Hypotheses in a Given Model 202
19.4.2 Tests of a Given Null Hypothesis Against Different Models 205
19.5 Exercises 210
19.6 References 211

20 Set Estimation and Confidence Regions 213
20.1 Definitions 213
20.2 Examples 214
20.2.1 Interval Estimation of the Mean of a Normal Distribution With Known Variance 214
20.2.2 Interval Estimation of the Mean of a Normal Distribution With Unknown Variance 215
20.2.3 Confidence Region for the Mean and the Variance of a Normal Distribution 215
20.2.4 Confidence Region for the Parameters of a Linear Model 217
20.3 Conservative Confidence Regions 218
20.4 Optimality of a Confidence Region 219
20.5 Examples of UMPU Confidence Regions 223
20.5.1 Interval Estimation of the Mean of a Normal Distribution With Unknown Variance 223
20.5.2 Confidence Interval for the Parameters of a Linear Model 224
20.6 Asymptotic Confidence Regions 224

20.7 Examples of Asymptotic Confidence Regions 227
20.8 Bayesian Confidence Regions 228
20.9 Exercises 229
20.10 References 231

21 Inequality Constraints: Estimation and Testing 233
21.1 Some Examples 233
21.1.1 Monotony Constraints 234
21.1.2 Barycentric Interpretation of Some Parameters 234
21.1.3 Positive Definiteness of Symmetric Matrices 235
21.1.4 Stability Conditions 236
21.2 Estimation Under Inequality Constraints 237
21.2.1 Constrained Estimation of the Linear Model 237
21.2.2 General Results 243
21.3 $H_0 : \{g(\theta) = 0\}$ versus $H_1 : \{g(\theta) \geq 0$ and $g(\theta) \neq 0\}$ 247
21.3.1 The Main Test Statistics 247
21.3.2 Asymptotically Equivalent Optimization Problems 253
21.3.3 Asymptotic Distribution of the Test Statistics Under the Null Hypothesis H_0 255
21.3.4 Examples 261
21.4 Tests of $H_0 : \{g(\theta) \geq 0\}$ 266
21.4.1 Some General Remarks 266
21.4.2 A Special Case 267
21.4.3 An Example 271
21.5 Exercises 272
21.6 References 273

22 Nonnested Models 275
22.1 Kullback Proximity and Pseudo True Values 276
22.1.1 Definition 276
22.1.2 Pseudo True Values and Proximities 278
22.2 Tests of Nonnested Hypotheses 282
22.2.1 Nonnested Hypotheses 282
22.2.2 Generalized Wald Tests 283
22.2.3 Generalized Score Tests 287
22.2.4 Selection of Regressors in Linear Models 288
22.2.5 Choice Among Qualitative Variable Models 290
22.2.6 Cox Testing Procedure 293
22.2.7 Using a Nesting Model 297

22.3 Model Selection Criteria 304
22.3.1 Linear Models 304
22.3.2 The Akaike Criterion 307
22.3.3 Bayesian Criteria 312
22.4 Appendix 1 313
22.5 Appendix 2 317
22.6 Exercises 320
22.7 References 322

23 Asymptotic Efficiency 325
23.1 Asymptotic Comparison of Estimators 325
23.1.1 First-Order Asymptotic Efficiency 326
23.1.2 Higher-Order Expansions 334
23.1.3 Edgeworth Expansion 343
23.2 Asymptotic Comparison of Tests 345
23.2.1 Statement of the Problem and Various Solutions 345
23.2.2 Concepts 350
23.2.3 Bahadur's Relative Efficiency 354
23.2.4 Pitman's Relative Efficiency 362
23.2.5 Direct Approach Based on Local Alternatives 366
23.3 Semiparametric Efficiency 370
23.3.1 Information Contained in Identifying Conditions 370
23.3.2 Asymptotically Efficient Semiparametric Estimators 376
23.3.3 A Two-Step Method For Obtaining an Asymptotically Efficient Estimator 377
23.4 References 380

24 Asymptotic Theory 383
24.1 Existence of an Extremum Estimator 383
24.2 Consistency of an Extremum Estimator 386
24.2.1 Almost Sure Consistency 386
24.2.2 Some Consistency Results 387
24.2.3 Conditions for the Uniform Almost Sure Convergence of the Objective Function 388
24.2.4 Consistency of Quasi Generalized Extremum Estimators 392
24.3 First-Order Conditions 394
24.3.1 Unconstrained Extremum Estimators 394

24.3.2 Constrained Extremum Estimators 395
24.4 Taylor Series Expansions 397
24.4.1 Infinitely Small Quantities in Probability 398
24.4.2 Taylor Series Expansion of the First-Order Conditions and Asymptotic Normality 401
24.5 References 404

Review of Linear Algebra and Matrix Calculus 405
A.1 Linear Algebra and Notations 405
A.1.1 Vector Spaces 405
A.1.2 Linear Mappings 406
A.1.3 Bases 406
A.1.4 Supplementary Subspaces 407
A.1.5 Matrix Representations 408
A.1.6 Matrix Operations 409
A.1.7 Change of Bases 411
A.2 Some Complements on Matrices 413
A.2.1 Trace of a Matrix 413
A.2.2 Partitioned Matrices 414
A.3 A Review of Quadratic Forms 420
A.3.1 Definition and Characterization 420
A.3.2 Positive Quadratic Forms 421
A.3.3 Symmetric Matrix Representations of a Linear Mapping 423
A.3.4 Relationships among $\ker \mathbf{A}$, Im $\mathbf{A}$, $\ker \mathbf{A}'$, and Im $\mathbf{A}'$ 423
A.4 Ordering Symmetric Matrices 424
A.4.1 Definition of an Order Relation 424
A.4.2 Comparing Ordered Spectra 425
A.4.3 Square Root of a Positive Semidefinite Matrix 427
A.4.4 A Geometric Interpretation 429
A.5 Projections 430
A.5.1 Definition and Characterization 431
A.5.2 Matrix Representations of Projections 433
A.5.3 Some Properties of Orthogonal Projections 434
A.5.4 An Optimal Property of Orthogonal Projections 436
A.6 Generalized Inverses 437
A.6.1 Definition and Construction 437
A.6.2 Invariance Properties 439
A.6.3 Solving Linear Systems of Equations 441

Review of Probability 443
B.1 Review of Integration Theory 443
B.1.1 Random Experiments 443
B.1.2 Characterization of a Probability Distribution or a Measure 444
B.1.3 Integration with Respect to a Measure 446
B.1.4 Density Functions 447
B.1.5 Random Variables 448
B.1.6 Properties of the Expectation of a Real Random Variable 449
B.1.7 Some Descriptive Measures of the Probability Distribution of a Real Random Variable 450
B.1.8 Some Probability Distributions on $\mathbb{R}$ 450
B.1.9 Random Vectors and Matrices 451
B.2 Independence and Conditioning 458
B.2.1 Conditional Probability Distributions 458
B.2.2 Conditional Expectations 461
B.2.3 Properties of the Conditional Expectation 462
B.2.4 Linear Regression 465
B.2.5 The Partial Variance Covariance Matrix 468
B.2.6 Independence 470
B.2.7 Conditional Independence 473
B.3 Normal Distributions 475
B.3.1 Univariate Normal Distributions 475
B.3.2 Multivariate Normal Distributions 477
B.3.3 Independence and Conditioning 480
B.3.4 Truncated Normal Distributions 483
B.4 Distributions Derived from the Normal Distribution 487
B.4.1 Chi-Square Distributions 487
B.4.2 Quadratic Forms of Gaussian Vectors 490
B.4.3 Student Distributions 492
B.4.4 Fisher Distributions 495
B.5 Stochastic Processes 496
B.5.1 Definition 496
B.5.2 Types of Convergence of Random Vectors 497
B.5.3 An Example of Stationary Processes: The Autoregressive Process of Order 1 499
B.5.4 Convergence in Distribution 501
B.5.5 Poisson Processes 502
B.6 Asymptotic Theorems 504

B.6.1 Laws of Large Numbers 504
B.6.2 Central Limit Theorems 506
B.6.3 Applications to Continuous Differentiable Functions 509
B.6.4 Other Asymptotic Theorems 510

CHAPTER 14

Introduction to Tests of Hypotheses

14.1 Testing Theory and Modelling

The purpose of the methods described in the first volume is to specify and estimate, on the basis of the data, a model of which the validity is not questioned. On the contrary, in the theory of hypothesis testing, which is the general topic of the second volume the validity of the model is now challenged.

For instance, one may wonder whether the specified model is not too "large," i.e., whether a submodel defined by a subset of the family $\mathcal{P}$ of possible probability distributions is not preferable. This is the basis of *significance tests.* Conversely, one may wonder whether the specified model is not too restrictive, i.e., whether the true distribution that has generated the observations actually belongs to $\mathcal{P}$. In the latter case, one frequently talks about *specification tests.* As a matter of fact, we shall not make a distinction between these two kinds of testing situations, for the approach that is generally considered in specification testing is to nest $\mathcal{P}$ in a larger family and to examine whether $\mathcal{P}$ is an acceptable restriction of this larger family. Hence the second problem reduces to the first problem. There exists, however, another approach to the problem of specification testing. This will be discussed in Chapter 22 when studying *nonnested hypotheses tests.*

Over the last fifty years, the statistical methods of the theory of hypothesis testing have considerably developed under the impulse of statisticians such as J. Neyman, E. Lehman and A. Wald. As for the theory of statistics in general, this development has its source in the increasing role of probabilistic modelling as a scientific tool. Another reason, however, which may be more fundamental and specific to the

theory of hypothesis testing, is the progressive disappearence of the idea that a model can be validated with certainty on the basis of the data. Such an idea, which was frequently held during the nineteen century, was gradually forsaken especially at the beginning of the twentieth century when physicists started to question the theory of classical mechanics considered up to then as the definitive theory. The relative value of a model in a collection of competing models then proved to be a valuable concept. As a natural consequence, the theory of hypothesis testing, whose main purpose is to arbitrate among models, received an increasing interest.

14.2 Hypotheses

A testing problem is defined by a statistical model $(\mathcal{Y}, \mathcal{P})$ and by a partitioning of the family $\mathcal{P}$ into two subfamilies $\mathcal{P}_0$ and $\mathcal{P}_1 = {}^c\mathcal{P}_0$. These two subfamilies define respectively two *hypotheses* about the true distribution P_0 generating the observations, namely

$$H_0 : P_0 \in \mathcal{P}_0$$

and

$$H_1 : P_0 \in \mathcal{P}_1.$$

It is frequently convenient to identify H_0 with $\mathcal{P}_0$ and H_1 with $\mathcal{P}_1$. Although the two hypotheses H_0 and H_1 play a symmetric role in this section, they are given hereafter two different names: H_0 is called the *null hypothesis* while H_1 is called the *alternative hypothesis.* The union of H_0 and H_1 defines the hypothesis $H : P_0 \in \mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$, which is called the *general* or *maintained hypothesis.*

Definition 14.1: *A hypothesis is called simple if it contains a unique probability distribution. It is called composite otherwise.*

In a parametric model $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$ the hypotheses H_0 and H_1 are defined, in general, by two subsets Θ_0 and $\Theta_1 = {}^c\Theta_0$ of Θ. When the model is identified, such a definition of the null and alternative hypotheses is identical to that based on a partition of $\mathcal{P}$ into $\mathcal{P}_0$ and $\mathcal{P}_1$. This is because the mapping that associates P_θ to θ is a one-to-one and onto mapping, i.e., a bijective mapping from Θ to $\mathcal{P}$. When the model is not identified, however, a difficulty arises. Specifically, there may exist values for the parameter $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ leading to the same probability distribution, i.e., such that $P_{\theta_0} = P_{\theta_1}$. In this case the subsets $\mathcal{P}_0 = \{P_\theta, \theta \in \Theta_0\}$ and $\mathcal{P}_1 = \{P_\theta, \theta \in \Theta_1\}$ are no longer disjoint.

Definition 14.2: *A testing problem defined by* Θ_0 *and* $\Theta_1 = {}^c\Theta_0$ *is identified if* P_{θ_0} *is different from* P_{θ_1} *for every* $\theta_0 \in \Theta_0$ *and* $\theta_1 \in \Theta_1$.

It is obvious that a model is identified if and only if every testing problem is identified. As Example 14.4 illustrates, however, *some* testing problems can be identified even though the model is not identified.

14.3 Examples

Example 14.1: A machine produces steel balls whose diameters are independently and identically distributed as $N(\theta, \sigma_0^2)$. It is assumed that the accuracy σ_0^2 of the machine is a characteristic known to the investigator and that the mean diameter θ of the produced steel balls is a parameter that can be chosen. One observes n diameters $Y_1, \ldots, Y_n$ and one wishes to test whether the tuning of the machine corresponds to the posted value θ_0.

In this example the statistical model is

$$\left(\mathbb{R}^n, \left\{(N(\theta, \sigma_0^2))^{\otimes n}, \theta \in \mathbb{R}^+\right\}\right).$$

The null hypothesis of good tuning is H_0: $\theta = \theta_0$ and the alternative hypothesis is H_1: $\theta \neq \theta_0$. The null hypothesis is simple while the alternative hypothesis is composite.

Example 14.2: It is assumed that the production Q_t of a given commodity at time t, $t = 1, \ldots, T$, can be modelled by the Cobb–Douglas production function

$$\log Q_t = a + b \log N_t + c \log K_t + u_t,$$

where N_t denotes the quantity of labor input and K_t denotes the quantity of capital input. It is assumed that the random disturbances u_t, $t = 1, \ldots, T$, are independently and identically distributed as $N(0, \sigma^2)$.

The model is parametric. If $\log Q_t$, $t = 1, \ldots, T$, are viewed as the observations, the model is

$$\left(\mathbb{R}^T, \left\{\bigotimes_{t=1}^{T} N(a + b \log N_t + c \log K_t, \sigma^2), (a, b, c, \sigma^2) \in \mathbb{R}^3 \times \mathbb{R}^+\right\}\right).$$

One may want to test the hypothesis H_0 of constant returns, i.e., the property that multiplying labor input and capital input by a same factor leads to multiplying production by this factor. Such a hypothesis

is identical to the condition that the Cobb–Douglas production function is homogenous of degree one. In terms of the parameters, this translates into the condition $b+c=1$. It is clear that the two hypotheses H_0 and H_1 are composite. For instance, H_0 is given by

$$\Theta_0 = \left\{(a,b,c,\sigma^2) \in \mathbb{R}^3 \times \mathbb{R}^+, b+c=1\right\}.$$

One may also want to question the hypothesis of independence among the u_t's. A method for dealing with such a "specification" testing situation is to nest the preceding model into a larger model where the disturbances u_t's satisfy the first-order autoregressive process

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad |\rho| < 1$$

and the ε_t's are independently and identically distributed as $N(0,\sigma_\varepsilon^2)$. The null hypothesis of independence of the u_t's is then characterized by the condition $\rho = 0$.

Example 14.3: A consumption survey provides observations on health expenditures and incomes of n households, (C_i, R_i), $i = 1,\ldots,n$. It is assumed that the pairs (C_i, R_i) are independently and identically distributed with unknown density $f(c_i, r_i)$ with respect to the Lebesgue measure λ_2^+ on $\mathbb{R}^{+2}$. The statistical model is defined by the family

$$\mathcal{P} = \left\{\prod_{i=1}^{n} f(c_i,r_i)\cdot\lambda_{2n}^+,\ f \text{ arbitrary on } \mathbb{R}^{+2}\right\}.$$

Suppose that one wishes to test whether health expenditures are independent of incomes. The null hypothesis corresponds to the subfamily

$$\mathcal{P}_0 = \left\{\prod_{i=1}^{n} g(c_i)h(r_i)\cdot\lambda_{2n}^+,\ g \text{ and } h \text{ arbitrary on } \mathbb{R}^+\right\}.$$

The model is nonparametric. The null and alternative hypotheses are composite.

Example 14.4: At time $t = 1,\ldots,T$, the quantity exchanged Q_t and the price p_t of an agricultural product are determined by the demand equation

$$Q_t = \alpha p_t + \beta x_{t-1} + \gamma z_{t-1} + \delta + u_t,$$

and the supply equation

$$Q_t = a p_t^* + b + v_t,$$

where x_{t-1} and z_{t-1} are some variables treated as nonstochastic, u_t and v_t are zero-mean random errors uncorrelated contemporaneously and over time. The variable p_t^* denotes the producers' expectation at time $t-1$ of price at time t. In addition, it is assumed that this expectation is a function of past exogenous variables and is given by

$$p_t^* = \phi_1 x_{t-1} + \phi_2 z_{t-1} + \phi_3.$$

Using this expression in the supply equation, one obtains

$$\begin{cases} Q_t = \alpha p_t + \beta x_{t-1} + \gamma z_{t-1} + \delta + u_t, \\ Q_t = \beta_1 x_{t-1} + \beta_2 z_{t-1} + \beta_3 + \nu_t, \end{cases}$$

where

$$\begin{aligned} \beta_1 &= a\phi_1, \\ \beta_2 &= a\phi_2, \\ \beta_3 &= b + a\phi_3. \end{aligned}$$

The parameters a, b, ϕ_1, ϕ_2, and ϕ_3 are not first-order identified although the parameters β_1, β_2, and β_3 are.

One wishes to test whether price expectations are "rational," i.e., whether they coincide with the optimal predictions $p_t^* = E_{t-1}p_t$, where $E_{t-1}p_t$ is the conditional expectation of p_t given the variables known at time $t-1$. Taking first the conditional expectation of the demand and supply equations and then, the difference between the resulting equations, one obtains

$$p_t^* = E_{t-1}p_t = \frac{\beta}{a-\alpha}x_{t-1} + \frac{\gamma}{a-\alpha}z_{t-1} + \frac{\delta - b}{a-\alpha}.$$

Thus the null hypothesis of "rational" expectations can be written as

$$H_0 : \phi_1 = \frac{\beta}{a-\alpha}, \quad \phi_2 = \frac{\gamma}{a-\alpha}, \quad \phi_3 = \frac{\delta - b}{a-\alpha}.$$

The hypothesis H_0 is not identified since ϕ_1, ϕ_2, and ϕ_3 are not identified.

The null hypothesis H_0, however, implies

$$H_0^* : \frac{\phi_1}{\phi_2} = \frac{\beta}{\gamma},$$

i.e.

$$H_0^* : \frac{\beta_1}{\beta_2} = \frac{\beta}{\gamma}.$$

Note that H_0^* is identified. Thus, although not identified, the hypothesis of rational expectations can be tested indirectly by testing one of its consequences.

Example 14.5: It is assumed that a price index Y_t can be described by

$$Y_t = a\exp(bt) + c + u_t, \quad t = 1, \ldots, T,$$

where a, b, and c are some unknown parameters, the u_t's are independent normal random variables with zero means and common variance σ^2. The model is parametric and given by

$$\left(\mathcal{Y} = I\!R^T, \mathcal{P} = \left\{\bigotimes_{t=1}^{T} N(a\exp(bt) + c, \sigma^2),\ (a, b, c, \sigma^2) \in I\!R^3 \times I\!R^{+*}\right\}\right).$$

The model is not identified. For instance, the same probability distribution is obtained from the following two sets of parameters $(a = 0, b = 1, c = 1, \sigma^2 = 1)$ and $(a = 0, b = 2, c = 1, \sigma^2 = 1)$.

Consider the null hypothesis of no trend. This hypothesis can be written in many ways. First, it can be written as $a = 0$, i.e., $\Theta_0 = \{(a, b, c, \sigma^2),\ a = 0\}$ or $\Theta_0 = \{0\} \times I\!R^2 \times I\!R^{+*}$. The two hypotheses H_0 and H_1 are composite. The testing problem thus defined is not identified because the two values $\theta_0 = (0, b_0, c_0, \sigma_0^2) \in \Theta_0$ and $\theta_1 = (a \neq 0, 0, c_0 - a, \sigma_0^2) \in \Theta_1$ lead to the same probability distribution.

Alternatively, the hypothesis of no trend can be written as H_0^*: $b = 0$. This does not define an identified testing problem because the values $\theta_0 = (0, 0, c_0, \sigma_0^2) \in \Theta_0$ and $\theta_1 = (0, b \neq 0, c_0, \sigma_0^2) \in \Theta_1$ lead to the same probability distribution. Similarly, the hypothesis H_0^{**}: $a = b = 0$ does not define an identified testing problem. However H_0^{***}: $ab = 0$ is identified.

14.4 Definition of a Test

In a testing problem a statistician must choose between two decisions, which are d_0: accept H_0 and d_1: accept H_1. Thus the set $\mathcal{D}$ of decisions is reduced to two elements, i.e., $\mathcal{D} = \{d_0, d_1\}$

Definition 14.3: *Given a model $(\mathcal{Y}, \mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1)$, a test is a decision rule with set of decisions $\mathcal{D} = \{d_0, d_1\}$ corresponding to $\mathcal{P}_0$ and $\mathcal{P}_1$.*

In particular, we have

Definition 14.4: *A nonrandomized test is a nonrandomized decision rule, i.e., a mapping δ from $\mathcal{Y}$ to $\mathcal{D} = \{d_0, d_1\}$.*

From Definition 14.4 it follows that a nonrandomized test δ is fully characterized by the inverse image $\delta^{-1}(d_1)$ of d_1, which is the set of observations $y \in \mathcal{Y}$ such that the hypothesis H_1 is accepted. This subset of $\mathcal{Y}$ is denoted W_δ, or more compactly, W. It is called the *critical region* of the test δ.

Allowing for randomized decision rules is, however, more important in the theory of hypothesis testing than in estimation theory. This is because the set $\mathcal{D} = \{d_0, d_1\}$ is not convex. Hence Property 5.1, which shows how to improve upon randomized decision rules using nonrandomized decision rules, no longer holds.

Definition 14.5: *A randomized test is a randomized decision rule, i.e., a mapping m from $\mathcal{Y}$ to the set of probability distributions on $\mathcal{D} = \{d_0, d_1\}$.*

A randomized test is fully characterized by a function $m_y(d_1)$ that associates a probability for accepting H_1 to each $y \in \mathcal{Y}$. The probability for accepting H_0 is then $m_y(d_0) = 1 - m_y(d_1)$. The function $m_y(d_1)$ is denoted $\phi_m(y)$ or $\phi(y)$, and is called the *test function* of the test m.

A nonrandomized test is a particular randomized test, where the decision d_1 is taken with probability one when y belongs to the critical region W and with probability zero otherwise. Thus the test function of a nonrandomized test is the indicator function $\phi = \mathbb{1}_W$. Conversely, if ϕ is an indicator function, the corresponding test is a nonrandomized test whose critical region is $W = \{y : \phi(y) = 1\}$.

In general, the test function ϕ is defined on $\mathcal{Y}$ with values in $[0, 1]$. In fact, there is a one-to-one and onto mapping between the set of randomized tests and the set of functions defined on $\mathcal{Y}$ with values in $[0, 1]$. As it can be checked, the latter set has the desirable property of being convex.

14.5 Types of Errors

To use standard notation, we assume hereafter that the model is parametric and that the hypotheses H_0 and H_1 are associated with sets Θ_0 and $\Theta_1 = \Theta - \Theta_0$. The following definitions, however, can be easily

extended to the case where the model is nonparametric or semiparametric.

In general, a statistician cannot take a correct decision with certainty on the basis of the outcome y of a random experiment. Specifically, two types of errors can be made:

- Choose d_1, i.e., H_1, while H_0 is true. This error is called a *Type I error.*
- Choose d_0, i.e., H_0, while H_1 is true. This error is called a *Type II error.*

This is summarized in the following table.

State / Decision	H_0 true	H_1 true
d_0	No error	Type II error
d_1	Type I error	No error

Consider now the probabilities of the two types of errors associated with a nonrandomized test (to begin with). A nonrandomized test is characterized by its critical region W. For a given value of the parameter θ, the probability of accepting H_0 is $P_\theta({}^cW)$ while the probability of accepting H_1 is $P_\theta(W)$. This leads to the following definition.

Definition 14.6: *The risk of Type I error of a nonrandomized test defined by a critical region W is the function $\alpha(W,\theta) = P_\theta(W)$ of θ defined on Θ_0.*

The risk of Type II error is the function $\beta(W,\theta) = P_\theta({}^cW) = 1 - P_\theta(W)$ of θ defined on Θ_1.

The function $P_\theta(W) = 1 - \beta(W,\theta) = \gamma(W,\theta)$ defined on Θ_1 is called the power function of the test W.

The risk functions of Type I and Type II errors are sometimes called *size functions.*

For a randomized test characterized by the test function ϕ, we know that the probability of accepting H_1 is $\phi(y)$ when the observation is y. Therefore, for a given value of the parameter θ, the probability of

accepting H_1 is

$$\int_{\mathcal{Y}} \phi(y) dP_\theta(y) = E_\theta \phi(Y) \equiv E_\theta \phi.$$

Similarly, the probability of accepting H_0 is $E_\theta(1-\phi) = 1 - E_\theta\phi$.

Definition 14.7: *For a randomized test ϕ, the risk of Type I error, the risk of Type II error, and the power function are given respectively by the functions*

$$\begin{array}{lcll} \alpha(\phi,\theta) & = & E_\theta\phi, & \theta \in \Theta_0, \\ \beta(\phi,\theta) & = & 1 - E_\theta\phi, & \theta \in \Theta_1, \\ \gamma(\phi,\theta) & = & E_\theta\phi, & \theta \in \Theta_1. \end{array}$$

14.6 Ordering Hypothesis Tests

The risk functions of Type I and Type II errors are related to the concept of risk functions in decision theory (see Chapter 2). To define such risk functions, it was first necessary to choose a loss function $L(d,\theta)$ measuring the loss incurred when the decision d is taken while the parameter value is θ. In Chapter 2, we argued that a natural choice for a loss function is one that is (i) equal to zero when a correct decision is made, and (ii) strictly positive when an incorrect decision is made. That is, a natural loss function satisfies

$$\begin{cases} L(d_1,\theta) > 0 \text{ and } L(d_0,\theta) = 0, & \forall\ \theta \in \Theta_0, \\ L(d_0,\theta) > 0 \text{ and } L(d_1,\theta) = 0, & \forall\ \theta \in \Theta_1. \end{cases}$$

From decision theory, the risk function associated with a randomized decision rule m is

$$R(m,\theta) = \int_{\mathcal{Y}} \left(\int_{\mathcal{D}} L(d,\theta) dm_y(d) \right) dP_\theta(y).$$

In the case of a randomized test m defined by the test function ϕ, the risk function becomes

$$R(\phi,\theta) = \begin{cases} L(d_1,\theta)E_\theta\phi & = L(d_1,\theta)\alpha(\phi,\theta), & \forall\ \theta \in \Theta_0, \\ L(d_0,\theta)E_\theta(1-\phi) & = L(d_0,\theta)\beta(\phi,\theta), & \forall\ \theta \in \Theta_1, \end{cases}$$

in view of the preceding form of the loss function.

The risk function $R(\phi, \theta)$ induces a preference relation on the set of hypothesis tests (see Definition 2.7). The induced ranking is readily characterized in terms of the risk functions α and β.

Property 14.1: *A test ϕ_1 is preferred to a test ϕ_2, which is denoted $\phi_1 \succeq \phi_2$, if and only if $\alpha(\phi_1, \theta) \leq \alpha(\phi_2, \theta)$, $\forall\ \theta \in \Theta_0$ and $\beta(\phi_1, \theta) \leq \beta(\phi_2, \theta)$, $\forall\ \theta \in \Theta_1$.*

PROOF: From Definition 2.7, the test ϕ_1 is preferred to the test ϕ_2 if and only if

$$R(\phi_1, \theta) \leq R(\phi_2, \theta), \quad \forall\ \theta \in \Theta.$$

Since

$$\begin{aligned} L(d_1, \theta) > 0, \qquad & \forall\ \theta \in \Theta_0, \\ L(d_0, \theta) > 0, \qquad & \forall\ \theta \in \Theta_1, \end{aligned}$$

it follows from the form of $R(\phi, \theta)$ that $R(\phi_1, \theta) \leq R(\phi_2, \theta)$, $\forall\ \theta \in \Theta$, is equivalent to the inequalities appearing in the statement of Property 14.1. □

Thus, to compare two tests ϕ_1 and ϕ_2, it suffices to compare the risk functions of Type I and Type II errors. Note that such a comparison does not depend on the loss function.

As in the theory of estimation, there does not exist in general an optimal test with respect to the preorder defined above.

Property 14.2: *With the exceptions of degenerated cases, there does not exist an optimal test.*

PROOF: Suppose that there exists an optimal test ϕ^*. Then ϕ^* must be preferred to (i) the constant test $\phi_1 = 0$, which consists of always accepting H_0, and (ii) the constant test $\phi_2 = 1$, which consists of always rejecting H_0. From Property 14.1 we have

$$\begin{aligned} \alpha(\phi^*, \theta) &= 0 = \int_{\mathcal{Y}} \phi^* dP_\theta, && \forall\ \theta \in \Theta_0, \\ \beta(\phi^*, \theta) &= 0 = \int_{\mathcal{Y}} (1 - \phi^*) dP_\theta, && \forall\ \theta \in \Theta_1. \end{aligned}$$

Because the function ϕ^* takes its values in $[0, 1]$, these equalities imply that, for every $\theta_0 \in \Theta_0$ and every $\theta_1 \in \Theta_1$, there exist two subsets A_{θ_0} and A_{θ_1} of $\mathcal{Y}$ such that

$$P_{\theta_0}(A_{\theta_0}) = 1, \ P_{\theta_1}(A_{\theta_1}) = 1$$

and

$$\phi^*(y) = 0, \forall y \in A_{\theta_0}, \ \phi^*(y) = 1, \ \forall y \in A_{\theta_1}.$$

It follows that, for every $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$, the probability distributions P_{θ_0} and P_{θ_1} have their supports included in the two disjoint subsets

$$A_0 = \bigcup_{\theta \in \Theta_0} A_{\theta_0} \quad \text{and} \quad A_1 = \bigcup_{\theta \in \Theta_1} A_{\theta_1}.$$

This implies, however, that we can choose with certainty between H_0 and H_1. It suffices to take the decision d_0 when y belongs to A_0 and the decision d_1 when y belongs to A_1. □

Though the preorder defined above does not admit an optimal test, such a preorder can be useful for excluding tests that are not admissible (see Definition 2.8).

Example 14.6: Consider one toss of a coin. Let

$$Y = \begin{cases} 1, & \text{if tail is observed,} \\ 0, & \text{if head is observed.} \end{cases}$$

The coin may not be fair. Let p denote the probability of observing "tail." The value of p is unknown and belongs to $[0, 1]$.

We are interested in the hypothesis $H_0 : p \geq 1/2$. In this example it is straightforward to characterize all the admissible tests. Note that the random variable Y is distributed as a Bernoulli $B(1, p)$. Then the risk functions α and β are given by

$$\begin{cases} \alpha(\phi, p) &= p\phi(1) + (1-p)\phi(0), & p \geq 1/2, \\ \beta(\phi, p) &= 1 - p\phi(1) - (1-p)\phi(0), & p < 1/2. \end{cases}$$

Hence a test ϕ_1 is preferred to a test ϕ_2 if

$$p\phi_1(1) + (1-p)\phi_1(0) \leq p\phi_2(1) + (1-p)\phi_2(0), \quad \forall\, p \geq 1/2,$$

and

$$p\phi_1(1) + (1-p)\phi_1(0) \geq p\phi_2(1) + (1-p)\phi_2(0), \quad \forall\, p < 1/2.$$

Let p approach $1/2$ in the preceding two inequalitites. Then a necessary condition for the two tests to be comparable is

$$\phi_1(1) + \phi_1(0) = \phi_2(1) + \phi_2(0).$$

In other words, a given test ϕ_1 is not comparable to another test ϕ_2 that does not satisfy this equality. Therefore, ϕ_1 is admissible if it is preferred to any test ϕ_2 that satisfies the above equality.

Given a value for $\phi(1) + \phi(0) = 2a$, $0 \leq a \leq 1$, we have

$$\begin{cases} \alpha(\phi, p) &= (2p-1)\phi(1) + 2a(1-p), & p \geq 1/2, \\ \beta(\phi, p) &= (1-2p)\phi(1) + 1 - 2a(1-p), & p < 1/2. \end{cases}$$

Thus, to minimize these two expressions with respect to ϕ for every p, it suffices to minimize $\phi(1)$ subject to the constraint $\phi(1) + \phi(0) = 2a$, where $0 \leq \phi(0) \leq 1$. Two cases must be distinguished:

if $0 \leq a \leq 1/2$, then $\phi(1) = 0$, $\phi(0) = 2a$,

if $1/2 \leq a \leq 1$, then $\phi(1) = 2a - 1$, $\phi(0) = 1$.

In particular, it follows that admissible tests are not always nonrandomized. Moreover, in this example, there are four nonrandomized tests. Three of them are admissible. They are obtained when

$$a = 0, \ a = 1/2, \text{ and } a = 1.$$

Only the nonrandomized test defined by $\phi(1) = 1$ and $\phi(0) = 0$ is not admissible.

14.7 Choice of a Test

Since, in general, an optimal test does not exist, the problem of choosing a test remains. Such a problem is dicussed in this section.

First, recall that, if S is a sufficient statistic for θ, then we can restrict ourselves to tests ϕ that are functions of observations through S, i.e., to tests of the form $\phi(y) = \Psi(S(y))$ (see Theorem 3.2). Thus it suffices to consider the model induced by S, i.e., $(\mathcal{S} = S(\mathcal{Y}), \{P_\theta^s, \theta \in \Theta\})$, and test functions $\psi(s)$ in this model. In fact, such a property is straightforward to establish because any test ϕ can be associated to a test $\psi(S) = E(\phi \mid S)$, which is based on S and which is independent of θ since S is sufficient. The latter test is equivalent to ϕ since

$$E_\theta \psi(S) = E_\theta E(\phi | S) = E_\theta \phi, \quad \forall \ \theta.$$

Hence

$$\begin{cases} \alpha(\psi, \theta) &= \alpha(\phi, \theta), \quad \forall \ \theta \in \Theta_0, \\ \beta(\psi, \theta) &= \beta(\phi, \theta), \quad \forall \theta \in \Theta_1. \end{cases}$$

In general, such a transformation by S greatly simplifies the problem of choosing a test. It does not, however, restrict the possible forms of

risk functions. Thus, as in point estimation, it is necessary to introduce some additional statistical principles to select an appropriate test. In this chapter we shall consider only two general principles. These are the Bayesian principle and the Neyman principle.

14.7.1 Bayesian Principle

Within the Bayesian framework there exists a prior distribution Π on Θ. Thus, for a test ϕ, we can define the Bayesian risk as

$$R_\Pi(\phi) = \int_\Theta R(\phi, \theta) d\Pi(\theta),$$

which can be used to compare hypothesis tests.

Given the restrictions on the loss function (see Section 14.6), we have

$$R_\Pi(\phi) = \int_\Theta \left(\int_\mathcal{Y} \left(L(d_1, \theta) \mathbb{I}_{\Theta_0}(\theta) \phi(y) + L(d_0, \theta) \mathbb{I}_{\Theta_1}(\theta)(1 - \phi(y)) \right) dP_\theta(y) \right) d\Pi(\theta).$$

Let Π_y denote the posterior distribution. Let P denote the marginal or *predictive* distribution of Y. Interverting the order of integration, we obtain

$$\begin{aligned} R_\Pi(\phi) &= \int_\mathcal{Y} \left(\int_\Theta \left(L(d_1, \theta) \mathbb{I}_{\Theta_0}(\theta) \phi(y) + \right. \right. \\ &\qquad \left. \left. L(d_0, \theta) \mathbb{I}_{\Theta_1}(\theta)(1 - \phi(y)) \right) d\Pi_y(\theta) \right) dP(y) \\ &= cst + \\ &\qquad \int_\mathcal{Y} \phi(y) \left(\int_{\Theta_0} L(d_1, \theta) d\Pi_y(\theta) - \int_{\Theta_1} L(d_0, \theta) d\Pi_y(\theta) \right) dP(y). \end{aligned}$$

An optimal test in the Bayesian sense is a test minimizing $R_\Pi(\phi)$. A solution is the nonrandomized test ϕ^* given by

$$\phi^*(y) = \begin{cases} 0, & \text{if } \int_{\Theta_0} L(d_1, \theta) d\Pi_y(\theta) \geq \int_{\Theta_1} L(d_0, \theta) d\Pi_y(\theta), \\ 1, & \text{if } \int_{\Theta_0} L(d_1, \theta) d\Pi_y(\theta) < \int_{\Theta_1} L(d_0, \theta) d\Pi_y(\theta). \end{cases}$$

In other words, y belongs to the critical region of the optimal Bayesian nonrandomized test if the average posterior loss incurred when accepting H_0 incorrectly is larger than the average posterior loss incurred when accepting H_1 incorrectly.

Example 14.7: Suppose that

$$L(d_1, \theta) = a_0, \forall\ \theta \in \Theta_0, \text{ and } L(d_0, \theta) = a_1, \forall\ \theta \in \Theta_1.$$

Then an optimal Bayesian nonrandomized test is

$$\phi^*(y) = \begin{cases} 0, & \text{if } a_0 \Pi_y(\Theta_0) \geq a_1 \Pi_y(\Theta_1), \\ 1, & \text{if } a_0 \Pi_y(\Theta_0) < a_1 \Pi_y(\Theta_1). \end{cases}$$

The test reduces to the simple comparison of the posterior probability ratio for Θ_0 and Θ_1 to the loss ratio. In particular, if accepting erroneously H_0 leads to an infinite loss ($a_1 \to +\infty$) and if the other loss is bounded, then this optimal test is equivalent to accepting H_1 as soon as $\Pi_y(\Theta_1)$ is strictly positive.

14.7.2 Neyman Principle

Up to now, the hypotheses H_0 and H_1 have played a symmetric role. On the contrary, within the Neyman framework, an asymmetry is introduced between these two hypotheses. Specifically, one attempts to control for the risk of Type I error $\alpha(\phi, \theta)$. More precisely, one considers only tests ϕ whose risks of Type I error are uniformly smaller than a given level $\alpha \in [0, 1]$, i.e., such that

$$\sup_{\theta \in \Theta_0} \alpha(\phi, \theta) \leq \alpha.$$

The level α is called the *significance level* of these tests. Then, among those tests, one searches for a test, if it exits, that minimizes the risk of Type II error $\beta(\phi, \theta)$ for every $\theta \in \Theta_1$, or equivalently, a test that maximizes the power function. Such a test, which is called *uniformly most powerful*, seldom exists. Thus, it is frequently necessary to introduce additional principles, which are discussed in the next chapters. These are the *unbiasedness principle, invariance principle, asymptotic principle, likelihood principle,* etc.

The Neyman principle, which is used frequently hereafter, is worth discussing further. In particular, one may question the asymmetry introduced between the two hypotheses H_0 and H_1.

In some situations, such an asymmetry is natural. For instance, consider a problem of quality control of industrial parts. The hypotheses H_0 and H_1 are defined by the value of the proportion of parts that are defective relative to a given level p_0. In this example, the buyer and the seller may have different views. The seller may want to control the

probability of destroying a defective lot. Within the Neyman framework, the seller will take $p \leq p_0$ as the null hypothesis. On the contrary, the buyer may want to control for the probability of buying a defective lot. To control for such a risk, the buyer will take $p \geq p_0$ as the null hypothesis.

In other situations, especially in econometrics, the null hypothesis is frequently the hypothesis that is commonly asserted or easiest to work with. Hence one wants to control for the probability of rejecting either a model that is commonly used or a simple model that is in fact acceptable.

It is useful to note that the asymmetry thus introduced does not necessarily "favor" H_0. Consider, for instance, the statistical model

$$(\mathbb{R}, \{N(\theta, 1), \theta = 0 \text{ or } 5\}).$$

Let the null hypothesis H_0 be defined by $\Theta_0 = \{\theta_0\} = \{0\}$. Let the alternative hypothesis H_1 be defined by $\Theta_1 = \{\theta_1\} = \{5\}$. In the next chapter we shall see that an optimal test in the Neyman sense for this problem is a nonrandomized test whose critical region is $W = \{Y > c\}$, where c is such that $P_{\theta_0}(W) = \alpha$. If $\alpha = 5\%$, then $c = 1.64$ and the risk of Type II error is $\beta = P_{\theta_1}((-\infty, 1, 64))$, which is close to zero. Thus the probability of rejecting H_0, when it is true, is much larger than the probability of rejecting H_1, when the latter is true. Hence, H_0 is not favored relative to H_1.

14.8 Exercises

EXERCISE 14.1: Consider the model $(\mathbb{R}^n, N(\theta, 1)^{\otimes n})$. One wants to test H_0: $\theta = 0$ against H_1: $\theta = 1$. Show that one can consider only tests ϕ that depend on the observations through the sample mean.

EXERCISE 14.2: Consider the preceding exercise. Let $\alpha(\phi)$ and $\beta(\phi)$ denote the risks of Type I and Type II errors of a test ϕ. Let Π_0 and Π_1 denote the prior probabilities associated with H_0 and H_1. It is assumed that the loss function L satisfies $L(d_0, 0) = L(d_1, 1) = 0$ and $L(d_0, 1) = L(d_1, 0) = 1$. Compute the Bayesian risk of a test ϕ. Find a test ϕ that minimizes the Bayesian risk.

EXERCISE 14.3: Consider the model $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$, where $\Theta = (\theta_0, \theta_1)$ and the P_θ's are discrete distributions. One wants to test H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$. Consider the Neyman framework. Show that one must include in the critical region points y for which the ratio

$P_{\theta_1}(y)/P_{\theta_0}(y)$ is highest. Does this procedure lead necessarily to a critical region whose probability under H_0 is equal to the given significance level α ?

EXERCISE 14.4: Consider a coin that may be either fair (hypothesis H_0) or have "head" on each of its sides (hypothesis H_1). The coin is tossed once. The loss incurred is zero if one makes the right decision and is equal to one otherwise. For a given test ϕ, compute the risks of Type I and Type II errors. Find the test that minimizes the maximum of these two risks. Such a test is called a *minimax* test.

14.9 References

Barra, J.R. (1970). *Notions fondamentales de statistique mathématique*, Dunod.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, Wiley.

Neyman, J. and Pearson, E. (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Phil. Trans. Roy. Soc.*, Ser. A, 231, 289–337.

Silvey, S.D. (1970). *Statistical Inference*, Chapman and Hall.

CHAPTER 15

Uniformly Most Powerful Tests

15.1 Tests of a Simple Hypothesis Against a Simple Hypothesis

The simplest testing problem arises when the null and alternative hypotheses are both simple. Then the family $\mathcal{P}$ of the statistical model $(\mathcal{Y}, \mathcal{P})$, consists of two probability distributions P_0 and P_1. The null hypothesis H_0 corresponds to the probability distribution P_0 while H_1 corresponds to P_1. Though this type of problem is seldom encountered in practice, its study will allow us to tackle more complex situations in subsequent sections.

15.1.1 Risk Diagram

When the two hypotheses H_0 and H_1 are simple, the risk functions of Type I and Type II errors associated with a test ϕ are reduced to the two real numbers

$$\alpha(\phi) = E_0\phi \quad \text{and} \quad \beta(\phi) = 1 - E_1\phi,$$

where E_0 and E_1 denote the mathematical expectations with respect to P_0 and P_1 respectively. Thus, to every test ϕ, we can associate a point $M(\phi)$ with coordinates $\big(\alpha(\phi), \beta(\phi)\big)$ in $\mathbb{R}^2$.

Definition 15.1: *A risk diagram is the set of points $M(\phi)$ when ϕ varies in the set of all possible test functions.*

The general form of a risk diagram is displayed in Figure 15.1. This follows from the main properties of a risk diagram, which are established next.

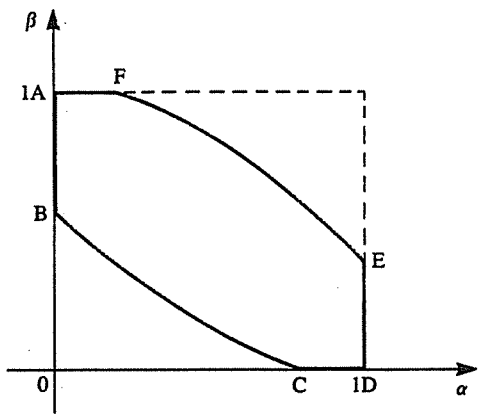


Figure 15.1: A Risk Diagram

Property 15.1: *The risk diagram is*

(i) included in the unit square $[0,1] \times [0,1]$,

(ii) symmetric about the point $(1/2, 1/2)$,

(iii) convex,

(iv) closed.

PROOF:

(i) The fact that the risk diagram is included in the unit square $[0,1] \times [0,1]$ is a consequence of the inequalities $0 \leq \phi(y) \leq 1, \forall\, y \in \mathcal{Y}$. Note that the vertices $(0,1)$ and $(1,0)$ always belong to the risk diagram since they correspond respectively to the nonrandomized tests $\phi = 0$ and $\phi = 1$.

(ii) If ϕ is a test function, then $1-\phi$ is also a test function. Because $M(1-\phi)$ is the point with coordinates $(1-\alpha(\phi), 1-\beta(\phi))$, then the risk diagram is symmetric about $(1/2, 1/2)$.

(iii) If ϕ_1 and ϕ_2 are two test functions, then for any λ satisfying $0 \leq \lambda \leq 1$, $\lambda\phi_1 + (1-\lambda)\phi_2$ is also a test function. Thus the risk diagram is convex.

(iv) The closedness of the risk diagram is more difficult to prove. It is a consequence of the weak compactness of the unit ball in the dual space. This implies that, if $\{\phi_n, n \in \mathbf{N}\}$ is a sequence of tests such that

$\alpha(\phi_n)$ and $\beta(\phi_n)$ converge to some limits α_∞ and β_∞ respectively, then there exists a test ϕ_∞ such that $\alpha_\infty = \alpha(\phi_\infty)$ and $\beta_\infty = \beta(\phi_\infty)$. □

Example 15.1:

a) Consider a random experiment which consists of drawing with replacement two balls from an urn containing many white balls and black balls in proportions p and $1-p$ respectively. Suppose that the parameter p can take only two values, namely $p = 1/3$ and $p = 1/2$. One wants to test the null hypothesis H_0: $p = 1/3$ against the alternative hypothesis H_1: $p = 1/2$. Let "0" denote the drawing of a black ball and "1" denote the drawing of a white ball. The set of possible outcomes consists of four elements, which are $(0,0), (0,1), (1,0)$, and $(1,1)$.

A test ϕ is defined by the four real numbers $\phi(0,0), \phi(0,1), \phi(1,0)$, and $\phi(1,1)$. The risks of Type I and Type II errors associated with ϕ are

$$\begin{aligned} \alpha(\phi) &= \frac{4}{9}\phi(0,0) + \frac{2}{9}(\phi(0,1) + \phi(1,0)) + \frac{1}{9}\phi(1,1), \\ \beta(\phi) &= 1 - \frac{1}{4}(\phi(0,0) + \phi(0,1) + \phi(1,0) + \phi(1,1)). \end{aligned}$$

Thus the point $M(\phi)$ is defined by

$$\begin{aligned} M(\phi) = &\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \phi(0,0)\begin{pmatrix} +4/9 \\ -1/4 \end{pmatrix} \\ &+(\phi(0,1) + \phi(1,0))\begin{pmatrix} +2/9 \\ -1/4 \end{pmatrix} + \phi(1,1)\begin{pmatrix} +1/9 \\ -1/4 \end{pmatrix}. \end{aligned}$$

When the values $\phi(i,j)$, $i = 0,1$, $j = 0,1$, vary between 0 and 1, it is easy to see that the point $M(\phi)$ belongs to a set whose boundary is a polygonal with vertices A, B, C, D, E, and F. See Figure 15.2.

The vertices of this polygonal correspond to the nonrandomized tests given in the table following Figure 15.2.

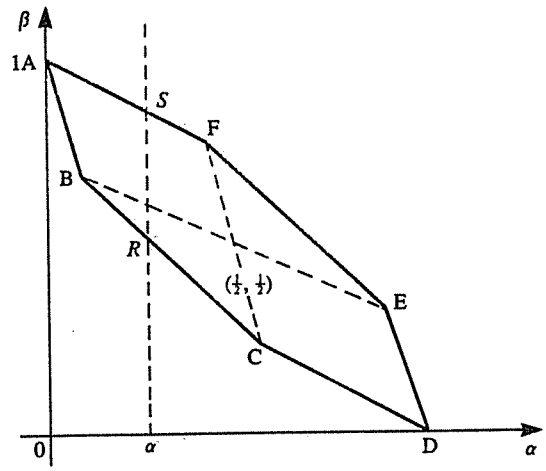


Figure 15.2: Risk Diagram

	$\phi(0,0)$	$\phi(0,1)$	$\phi(1,0)$	$\phi(1,1)$
A	0	0	0	0
B	0	0	0	1
C	0	1	1	1
D	1	1	1	1
E	1	1	1	0
F	1	0	0	0

b) In this example the admissible tests correspond to the points $M(\phi)$ such that the southwest orthant with origin $M(\phi)$ intersects the risk diagram at $M(\phi)$ only. Thus the points of the risk diagram that correspond to admissible tests are on the edges ABCD. See Figure 15.3.

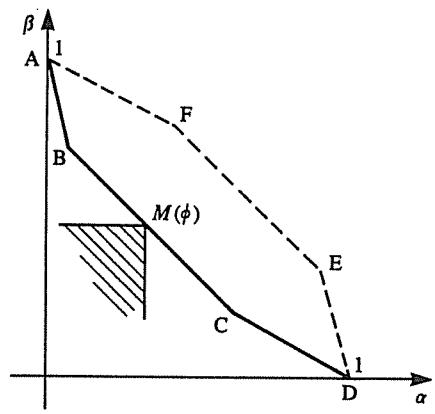


Figure 15.3: Admissible Tests

c) Now consider the Bayesian framework. Within such a framework, a prior distribution $\Pi = (\pi_0, \pi_1)$ is given. Thus the Bayesian risk is

$$R_\pi(\phi) = \pi_0 L(d_1, P_0)\alpha(\phi) + \pi_1 L(d_0, P_1)\beta(\phi).$$

Hence the tests ϕ whose Bayesian risks are equal to a given level r are associated to the points $M(\phi)$ that belong to the straight line defined by the equation

$$\pi_0 L(d_1, P_0)\alpha + \pi_1 L(d_0, P_1)\beta = r.$$

Therefore, to determine an optimal test in the Bayesian sense, it suffices to find the lowest straight line with slope $-\pi_0 L(d_1, P_0)/\pi_1 L(d_0, P_1)$ and having a nonempty intersection with the risk diagram.

When the prior distribution Π changes, it is easy to see that the set of optimal tests in the Bayesian sense is identical to the set of admissible tests (see Property 2.2). Moreover, given any Π, there always exists a nonrandomized test that is optimal in the Bayesian sense. This test corresponds to one of the vertices A, B, C, or D. See Figure 15.4.

d) Within the Neyman framework one only considers tests that have a risk of Type I error that is smaller than a given significance level

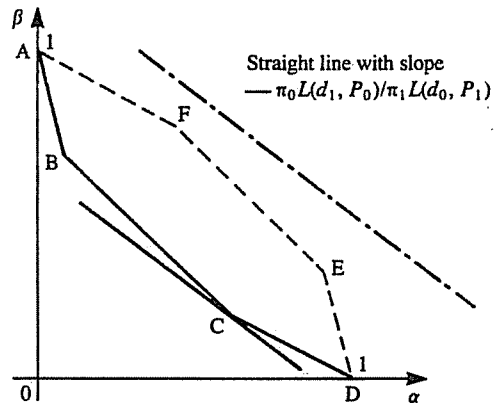


Figure 15.4: Bayesian Tests

α. The corresponding points in the risk diagram are those points that belong to the quadrilateral ABRS in Figure 15.2. Thus an optimal test in the Neyman sense corresponds to a point of this quadrilateral whose ordinate $\beta(\phi)$ is minimal. This point is R.

When the significance level α changes, it is easy to see that the set of optimal tests in the Neyman sense is, in this example, identical to the set of admissible tests. Note, however, that this property is not always true. For instance, in a risk diagram of the form given in Figure 15.1, the tests that correspond to the segment (C,D) are optimal in the Neyman sense (given an appropriate α) but are not admissible.

15.1.2 Neyman–Pearson Theorem

Given a significance level α, we now characterize the tests that are optimal in the Neyman sense. These tests are those that minimize the risk of Type II error $\beta(\phi)$, i.e., that maximize the power $E_1\phi$ subject to the constraint $E_0\phi \leq \alpha$.

Because there are only two probability distributions in the model, the model is always dominated by some measure μ (for instance, by P_0+P_1). Thus P_0 and P_1 admit some densities with respect to μ. These densities are denoted $\ell_0(y)$ and $\ell_1(y)$ respectively.

Definition 15.2: *A Neyman test is a test satisfying*

$$\begin{cases} \phi(y) = 1, & \text{if } \ell_1(y) > k\ell_0(y), \\ \phi(y) = 0, & \text{if } \ell_1(y) < k\ell_0(y), \end{cases}$$

where k is a real positive number.

Note that Definition 15.2 does not impose any condition on the test function ϕ when the equality $\ell_1(y) = k\ell_0(y)$ is satisfied.

Neyman tests have a natural interpretation. Namely, a hypothesis is accepted when it is sufficienty likely relative to the other hypothesis.

Remark 15.1: The test function of a Neyman test can be written as

$$\begin{cases} \phi(y) = 1, & \text{if } \log(\ell_1(y)/\ell_0(y)) = \log \ell_1(y) - \log \ell_0(y) > K, \\ \phi(y) = 0, & \text{if } \log(\ell_1(y)/\ell_0(y)) = \log \ell_1(y) - \log \ell_0(y) < K, \end{cases}$$

where $K = \log k$. Thus a Neyman test can be expressed in terms of the log-likelihoods, which are typically simpler than the likelihoods of the hypotheses.

Neyman–Pearson Theorem 15.1:

(i) For any α, $0 < \alpha < 1$, there exists a Neyman test ϕ such that $E_0\phi = \alpha$. Moreover, for those y's satisfying $\ell_1(y) = k\ell_0(y)$, where k is the real number defining ϕ, one can set $\phi = \gamma$ (independently of y).

(ii) For any α, $0 < \alpha < 1$, a Neyman test ϕ satisfying $E_0\phi = \alpha$ is optimal at the significance level α.

(iii) Conversely, for any given significance level α, $0 < \alpha < 1$, an optimal test is necessarily a Neyman test.

PROOF:

(i) Let ϕ be a Neyman test corresponding to a real number k (see Definition 15.4). Let $\phi(y) = \gamma$ if $\ell_1(y) = k\ell_0(y)$. Then we have

$$\begin{aligned} E_0\phi &= P_0(\ell_1(Y) > k\ell_0(Y)) + \gamma P_0(\ell_1(Y) = k\ell_0(Y)) \\ &= P_0\left(\frac{\ell_1(Y)}{\ell_0(Y)} > k\right) + \gamma P_0\left(\frac{\ell_1(Y)}{\ell_0(Y)} = k\right). \end{aligned}$$

Let F denote the cumulative distribution of the likelihood ratio, i.e., of $\ell_1(Y)/\ell_0(Y)$. We have

$$F(k) = P_0\left(\frac{\ell_1(Y)}{\ell_0(Y)} \leq k\right).$$

Then $E_0\phi$ can be written as

$$E_0\phi = 1 - F(k) + \gamma(F(k) - F(k^-)),$$

where $F(k^-) = \lim_{x \uparrow k} F(x)$.

Since F is right continuous, increasing, and such that

$$F(0^-) = 0 \text{ and } \lim_{k \to +\infty} F(k) = 1,$$

then there exists a real number k_0 satisfying $F(k_0^-) \leq 1 - \alpha \leq F(k_0)$.

Two cases can be distinguished:

a) If $1 - \alpha = F(k_0)$, then the Neyman test ϕ corresponding to $k = k_0$ and $\gamma = 0$ satisfies $E_0\phi = \alpha$.

b) If $F(k_0^-) \leq 1 - \alpha < F(k_0)$, then the Neyman test ϕ corresponding to the constants $k = k_0$ and

$$\gamma = \frac{F(k_0) - (1 - \alpha)}{F(k_0) - F(k_0^-)}$$

satisfies $E_0\phi = \alpha$.

(ii) We now establish the optimality of Neyman tests. Let ϕ be a Neyman test such that $E_0\phi = \alpha$. Let k be the corresponding number. Let ϕ^* be another test satisfying the constraint $E_0\phi^* \leq \alpha$ on the risk of Type I error.

Then the integral

$$\int (\phi(y) - \phi^*(y)) (\ell_1(y) - k\ell_0(y)) \, d\mu(y)$$

is nonnegative. This follows from $\phi = 1 \geq \phi^*$ if $\ell_1(y) > k\ell_0(y)$ and $\phi = 0 \leq \phi^*$ if $\ell_1(y) < k\ell_0(y)$. Thus we have the inequalities

$$E_1(\phi - \phi^*) \geq kE_0(\phi - \phi^*) = k(\alpha - E_0\phi^*).$$

Because the risk of Type I error associated with ϕ^* is less or equal to α, it follows that $E_1\phi$ is greater or equal to $E_1\phi^*$. Thus ϕ is optimal at the significance level α.

(iii) It remains to prove the converse of the preceding result. Let $\tilde{\phi}$ be an optimal test with significance level α. Let ϕ be a Neyman test such that $E_0\phi = \alpha$. From (ii), ϕ is also optimal at level α. Because both tests are optimal, their power must be equal. Hence we have $E_1\tilde{\phi} = E_1\phi$.

Consider now the integral

$$\int \left(\phi(y) - \tilde{\phi}(y)\right) (\ell_1(y) - k\ell_0(y))\, d\mu(y),$$

where k is an arbitrary positive real number. This integral is equal to

$$E_1\phi - E_1\tilde{\phi} - k(E_0\phi - E_0\tilde{\phi}) = k(E_0\tilde{\phi} - \alpha) \leq 0.$$

Next choose k as the number corresponding to ϕ. From (ii) we know that the above integral is nonnegative. Hence this integral must be equal to zero. Because the integrand is nonnegative (see (ii)), this integrand must be equal to zero. This implies

$$\begin{cases} \tilde{\phi} = 1, \text{ if } \ell_1(y) > k\ell_0(y), \\ \tilde{\phi} = 0, \text{ if } \ell_1(y) < k\ell_0(y). \end{cases}$$

Thus $\tilde{\phi}$ is a Neyman test as required. □

Remark 15.2: If the probability distribution of the likelihood ratio $\ell_1(Y)/\ell_0(Y)$ is continuous under the null hypothesis H_0, then we have $P_0(\ell_1(Y) = k\ell_0(Y)) = 0$ for every k. Thus, for any given significance level α, there exists an optimal nonrandomized test whose critical region is $W = \{\ell_1(y) > k\ell_0(y)\}$, where k is defined by $P_0(\ell_1(Y) > k\ell_0(Y)) = \alpha$.

Remark 15.3: Suppose that there exists a sufficient statistic S. From the Factorization Criterion 3.5, we have

$$\begin{aligned} \ell_0(y) &= \psi_0(S(y))\lambda(y), \\ \ell_1(y) &= \psi_1(S(y))\lambda(y). \end{aligned}$$

Thus

$$\ell_1(y)/\ell_0(y) = \psi_1(S(y))/\psi_0(S(y)).$$

Hence an optimal Neyman test at the significance level α depends on the observations through S only.

Property 15.2: *The power $E_1\phi$ of an optimal test at the significance level α is necessarily greater or equal to α.*

PROOF: The optimal test must be preferred to the constant test equal to α, namely, to $\phi^* = \alpha$. Hence $E_1\phi \geq E_1\phi^* = \alpha$. □

15.1.3 Application to Exponential Families

We consider a one-parameter exponential model with densities

$$\ell(y;\theta) = C(\theta)h(y)\exp(Q(\theta)T(y)),$$

where the parameter θ can take only two values, i.e., $\theta \in \Theta = \{\theta_0, \theta_1\}$, where $\theta_1 > \theta_0$. We assume that the model is identified. Hence $Q(\theta_0)$ is different from $Q(\theta_1)$. Thus, without loss of generality, we can choose the function Q so that $Q(\theta_1) > Q(\theta_0)$.

We want to test the null hypothesis H_0: $\theta = \theta_0$ against the alternative hypothesis H_1: $\theta = \theta_1$. From Neyman–Pearson Theorem, there exists an optimal test. It is of the form

$$\begin{cases} \phi(y) = 1, \text{ if } \ell(y;\theta_1) > k\ell(y;\theta_0), \\ \phi(y) = \gamma, \text{ if } \ell(y;\theta_1) = k\ell(y;\theta_0), \\ \phi(y) = 0, \text{ if } \ell(y;\theta_1) < k\ell(y;\theta_0). \end{cases}$$

Since

$$\frac{\ell(y;\theta_1)}{\ell(y;\theta_0)} = \frac{C(\theta_1)}{C(\theta_0)}\exp\left((Q(\theta_1) - Q(\theta_0))T(y)\right),$$

and since $Q(\theta_1) > Q(\theta_0)$, this test can also be written as

$$\begin{cases} \phi(y) = 1, \text{ if } T(y) > c, \\ \phi(y) = \gamma, \text{ if } T(y) = c, \\ \phi(y) = 0, \text{ if } T(y) < c. \end{cases}$$

That is, the optimal Neyman test can be expressed as a simple function of the natural or canonical statistic T.

The constants γ and c are determined by the condition $E_0\phi = \alpha$. If $\theta_0 > \theta_1$, it is clear that a similar result holds by reversing the inequalities in the definition of ϕ given above.

Example 15.2: Consider a random sample of size n drawn from the normal distribution $N(\theta, 1)$. The model is exponential with densities

$$\ell(y;\theta) = (2\pi)^{-n/2}\exp\left(\frac{-n\theta^2}{2}\right)\exp\left(\frac{-\sum_{i=1}^n y_i^2}{2}\right)\exp(n\theta\bar{y}).$$

Thus $Q(\theta) = n\theta$ and $T(y) = \bar{y}$. Therefore, since the distribution of $\bar{Y}$ is continuous (for every θ), there exists an optimal nonrandomized test for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$, where $\theta_1 > \theta_0$.

This optimal test is defined by the critical region $W = \{\bar{y} > c\}$. The number c is determined by the condition $P_0(W) = \alpha$. Since $\bar{Y}$ is distributed as $N(\theta_0, 1/n)$ under H_0, we have

$$\alpha = P_0(\bar{Y} > c) = P_0(\sqrt{n}(\bar{Y} - \theta_0) > \sqrt{n}(c - \theta_0)) = 1 - \Phi(\sqrt{n}(c - \theta_0)),$$

where Φ is the cumulative function of the standard normal distribution. Hence

$$\sqrt{n}(c - \theta_0) = \Phi^{-1}(1 - \alpha),$$

which gives

$$c = \theta_0 + \frac{1}{\sqrt{n}}\Phi^{-1}(1 - \alpha).$$

Thus the optimal critical region is

$$W = \left\{\bar{y} > \theta_0 + \frac{1}{\sqrt{n}}\Phi^{-1}(1 - \alpha)\right\}.$$

Example 15.3: Consider two Poisson distributions $P_0 = \mathcal{P}(5)$ and $P_1 = \mathcal{P}(10)$. Suppose that there is only one observation y. We want to test H_0: $P = P_0$ against H_1: $P = P_1$ at the significance level $\alpha = 5\%$.

The model is exponential. The ratio

$$\frac{\ell_1(y)}{\ell_0(y)} = \frac{e^{-10}10^y y!}{e^{-5}5^y y!} = e^{-5}2^y$$

is a function of y. Thus an optimal test is of the form

$$\begin{cases} \phi = 1, \text{ if } y > c, \\ \phi = \gamma, \text{ if } y = c, \\ \phi = 0, \text{ if } y < c. \end{cases}$$

The numbers γ and c are determined by the condition

$$P_0(Y > c) + \gamma P_0(Y = c) = 0.05.$$

From the table of the Poisson distribution $\mathcal{P}(5)$, we have

$$P_0(Y > 9) = 0.032 \quad \text{and} \quad P_0(Y > 8) = 0.068.$$

Thus $c = 9$ and γ is determined by

$$0.032 + \gamma(0.068 - 0.032) = 0.05.$$

Hence $\gamma = 1/2$.

To summarize, the decision rule associated with the above optimal test is (i) to reject the null hypothesis H_0 if $y > 9$, (ii) to choose randomly between H_0 and H_1 with equal probabilities if $y = 9$, and (iii) to accept H_0 if $y < 9$.

If one insists on having a nonrandomized test, then it is necessary to change the risk of Type I error. Specifically, in this example, the risks closest to 5% leading to a nonrandomized test are $\alpha_1 = 3.2\%$ and $\alpha_2 = 6.8\%$. The corresponding nonrandomized tests are defined by the critical regions $W_1 = \{y > 9\}$ and $W_2 = \{y > 8\}$ respectively.

15.2 Tests of One-Sided Hypotheses

We consider a one-parameter model $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$, where Θ is an interval of $\mathbb{R}$. We assume that the model is dominated by a measure μ. Let $\{\ell(y;\theta), \theta \in \Theta\}$ denote the family of the densities of the distributions $P_\theta, \theta \in \Theta$. These densities are assumed to be strictly positive. We are interested in testing the null hypothesis H_0 defined by $\Theta_0 = \{\theta : \theta \leq \theta_0\}$ against the alternative hypothesis H_1 defined by $\Theta_1 = \{\theta : \theta > \theta_0\}$, where θ_0 is an element of Θ such that Θ_0 and Θ_1 are nonempty.

15.2.1 Monotone Likelihood Ratio Families

Definition 15.3: *The family $\{P_\theta, \theta \in \Theta\}$ is said to be a monotone likelihood ratio family if there exists a statistic $U(y)$ with values in $\mathbb{R}$ such that $\forall\ \theta' < \theta''$, the ratio $\ell(y;\theta'')/\ell(y;\theta')$ is a strictly increasing or decreasing function of U.*

By changing U to $-U$, there is no loss of generality in assuming that the ratios $\ell(y;\theta'')/\ell(y,\theta')$ for $\theta' < \theta''$ are strictly increasing functions of $U(y)$.

The next theorem due to Lehman states that there exists a uniformly most powerful (UMP) test under the conditions of Definition 15.3.

Property 15.3: *Suppose that $\{\ell(y;\theta), \theta \in \Theta\}$ is a strictly increasing likelihood ratio family. Then, for any α, $0 < \alpha < 1$, there exists a UMP test with significance level α for testing H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$. This test is given by*

$$\begin{cases} \phi = 1, \text{ if } U(y) > c, \\ \phi = \gamma, \text{ if } U(y) = c, \\ \phi = 0, \text{ if } U(y) < c, \end{cases}$$

where the constants c and γ are determined by the condition $E_{\theta_0}\phi = \alpha$.

PROOF: The proof consists of three steps.

a) Consider the problem of testing the simple hypothesis $\tilde{H}_0$: $\theta = \theta_0$ against the simple hypothesis $\tilde{H}_1$: $\theta = \theta''$, where θ'' is a given value greater than θ_0. From the Neyman–Pearson Theorem, an optimal test at level α is

$$\begin{cases} \phi = 1, \text{ if } \ell(y;\theta'') > k\ell(y;\theta_0), \\ \phi = \gamma, \text{ if } \ell(y;\theta'') = k\ell(y;\theta_0), \\ \phi = 0, \text{ if } \ell(y;\theta'') < k\ell(y;\theta_0), \end{cases}$$

From the monotone likelihood ratio property it follows that this test can be written as

$$\begin{cases} \phi = 1, \text{ if } U(y) > c, \\ \phi = \gamma, \text{ if } U(y) = c, \\ \phi = 0, \text{ if } U(y) < c, \end{cases}$$

where the constants c and γ are determined by $E_{\theta_0}\phi = \alpha$.

b) The preceding test does not depend on the choice of θ'', $\theta'' > \theta_0$. Therefore the test is UMP at level α for testing $\tilde{H}_0$: $\theta = \theta_0$ against H_1: $\theta > \theta_0$. This means that ϕ is uniformly more powerful than any test $\tilde{\phi}$ satisfying $E_{\theta_0}\tilde{\phi} \leq \alpha$. Hence ϕ is more powerful than any test ϕ^* satisfying $\sup_{\theta\in\Theta_0} E_\theta\phi^* \leq \alpha$, i.e., than any test with level less or equal to α.

c) To complete the proof, it is necessary to verify that the level of the test ϕ is also less or equal to α. Let $\theta' < \theta_0$. From the Neyman–Pearson Theorem, it follows that the test ϕ is optimal at level $E_{\theta'}\phi$ for testing H_0^*: $\theta = \theta'$ against H_1^*: $\theta = \theta_0$. From Property 15.2, it follows that

$$E_{\theta_0}\phi \geq E_{\theta'}\phi.$$

Hence $E_{\theta'}\phi \leq E_{\theta_0}\phi = \alpha, \forall\ \theta' < \theta_0$. Thus the level of ϕ is α. □

Remark 15.4: Suppose now that the null hypothesis H_0 is defined by $\theta \geq \theta_0$ and the alternative hypothesis H_1 is defined by $\theta < \theta_0$. Provided the conditions of Property 15.3 are satisfied, then a UMP test at level α is obtained by reversing the inequalities defining the UMP test given in this property.

In particular, if the probability distribution of U for $\theta = \theta_0$ is symmetric about a number a, then a UMP test at level α for testing H_0: $\theta \geq \theta_0$ against H_1: $\theta < \theta_0$ is

$$\begin{cases} \phi^*(y) = 1, \text{ if } U(y) < 2a - c, \\ \phi^*(y) = \gamma, \text{ if } U(y) = 2a - c, \\ \phi^*(y) = 0, \text{ if } U(y) > 2a - c, \end{cases}$$

where c and γ are the constants obtained for the test ϕ of Property 15.3. This follows from the symmetry of the probability distribution of U, which gives

$$\begin{aligned} P_{\theta_0}(U < 2a - c) + \gamma P_{\theta_0}(U = 2a - c) &= P_{\theta_0}(U > c) + \gamma P_{\theta_0}(U = c) \\ &= \alpha. \end{aligned}$$

15.2.2 Application to Exponential Families

We consider a one-parameter exponential model with densities

$$\ell(y;\theta) = C(\theta)h(y)\exp(Q(\theta)T(y)),$$

where $Q(\theta)$ is strictly increasing. The same reasoning as in Section 15.1.3 shows that such a family is an increasing likelihood ratio family in $U = T$. It follows that Property 15.3 applies to such a family. In the next table we give the form of the statistic U, which is used to construct UMP tests for testing H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$.

Distribution for one observation	U for one observation	U for a random sample of size n
$\mathcal{P}(\theta)$	Y	$\sum_{i=1}^n Y_i$ or $\bar{Y}$
$B(n,\theta)$	Y	$\sum_{i=1}^n Y_i$ or $\bar{Y}$
$N(\theta,\sigma_0^2)$ σ_0^2 known	Y	$\sum_{i=1}^n Y_i$ or $\bar{Y}$
$N(\mu,\theta)$ μ known	$(Y-\mu)^2$	$\sum_{i=1}^n (Y_i-\mu)^2$
$\gamma(\theta)$	$\log Y$	$\sum_{i=1}^n \log Y_i$ or $\prod_{i=1}^n Y_i$

In particular, the test in Example 15.2 is UMP at level α for testing the null hypothesis H_0: $\theta \leq \theta_0$ against the alternative hypothesis H_1: $\theta > \theta_0$. Similarly, the test in Example 15.3 is UMP for testing H_0: $\theta \leq 5$ against H_1: $\theta > 5$.

15.2.3 Locally Uniformly Most Powerful Tests and Score Tests

When the conditions of Property 15.3 are not satisfied, there may no longer exist a UMP test for testing H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$. In general, however, it is possible to construct a test that is *locally* more powerful for testing H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$. Such a test is optimal in the Neyman sense only in a neighborhood of the boundary value θ_0.

Definition 15.4: *Consider testing H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$.*

(i) A test ϕ is locally of level α if there exists a neighborhood ν of θ_0 such that

$$\sup_{\theta \leq \theta_0, \theta \in \nu} E_\theta \phi = E_{\theta_0} \phi = \alpha.$$

(ii) A test ϕ is locally more powerful than a test ϕ^* if there exists a neighborhood ν of θ_0 such that $E_\theta \phi > E_\theta \phi^*, \forall\ \theta > \theta_0, \theta \in \nu$.

(iii) A test ϕ is locally uniformly most powerful (LUMP) at level α if it is locally of level α and if it is locally more powerful than any other tests that are locally of level α.

Property 15.4: *Suppose that one considers only tests ϕ^* such that $E_\theta \phi^*$ is differentiable at θ_0. Then a test ϕ satisfying*

$$\left(\frac{\partial E_\theta \phi}{\partial \theta} \right)_{\theta_0} \neq 0$$

is LUMP at level α if and only if it is locally of level α and satisfies

$$\left(\frac{\partial E_\theta \phi}{\partial \theta} \right)_{\theta_0} \geq \left(\frac{\partial E_\theta \phi^*}{\partial \theta} \right)_{\theta_0},$$

for any other tests ϕ^* that are locally of level α.

PROOF: Property 15.4 follows from the Taylor expansion

$$\begin{aligned} E_\theta\phi - E_\theta\phi^* &= E_{\theta_0}\phi - E_{\theta_0}\phi^* \\ &\quad +(\theta-\theta_0)\left(\left(\frac{\partial E_\theta\phi}{\partial\theta}\right)_{\theta_0} - \left(\frac{\partial E_\theta\phi^*}{\partial\theta}\right)_{\theta_0}\right) + o(\theta-\theta_0) \\ &= (\theta-\theta_0)\left(\left(\frac{\partial E_\theta\phi}{\partial\theta}\right)_{\theta_0} - \left(\frac{\partial E_\theta\phi^*}{\partial\theta}\right)_{\theta_0}\right) + o(\theta-\theta_0). \end{aligned}$$

The desired result follows immediately. □

Property 15.5: *Suppose that it is possible to differentiate under the integral sign. Then any LUMP test at level α satisfies*

$$\begin{cases} \phi(y) = 1, \text{ if } \partial\ell(y;\theta_0)/\partial\theta > k\ell(y;\theta_0), \\ \phi(y) = 0, \text{ if } \partial\ell(y;\theta_0)/\partial\theta < k\ell(y;\theta_0). \end{cases}$$

In addition, one can set

$$\phi(y) = \gamma, \text{ if } \partial\ell(y;\theta_0)/\partial\theta = k\ell(y;\theta_0),$$

where k and γ are constants determined by the condition $E_{\theta_0}\phi = \alpha$.

PROOF: The proof is almost identical to the proof of the Neyman–Pearson Theorem. Note, however, that k is not necessarily positive. Nonetheless the proof still applies because we are now considering tests satisfying $E_{\theta_0}\phi^* = \alpha$ instead of $E_{\theta_0}\phi^* \leq \alpha$. □

Definition 15.5: *The LUMP test of Property 15.5 is called the score test because it is based on the score $S_{\theta_0}(Y) = \partial \log \ell(Y;\theta_0)/\partial\theta$.*

Remark 15.5: The score test can be viewed as a limit case of a Neyman test for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$, when θ_1 converges to θ_0 with $\theta_1 > \theta_0$. Specifically, from the Neyman–Pearson Theorem, a Neyman test is of the form

$$\phi(y) = 1, \text{ if } \frac{\log\ell(y;\theta_1) - \log\ell(y;\theta_0)}{\theta_1-\theta_0} > k,$$

$$\phi(y) = 0, \text{ if } \frac{\log\ell(y;\theta_1) - \log\ell(y;\theta_0)}{\theta_1-\theta_0} < k.$$

Then it suffices to note that the ratio $(\log\ell(y;\theta_1) - \log\ell(y;\theta_0))/(\theta_1-\theta_0)$ converges to $S_{\theta_0}(y) = \partial\log\ell(y;\theta_0)/\partial\theta$, when θ_1 converges to θ_0.

In the special case where the probability distribution of the score $\partial \log \ell(Y;\theta_0)/\partial\theta$ is continuous under the null hypothesis H_0, then there exists a nonrandomized LUMP test at level α. Its critical region is given by $\partial \log \ell(Y;\theta_0)/\partial\theta \geq k$, where k is defined by the condition $P_{\theta_0}(W) = \alpha$ on the risk of Type I error.

The exact distribution of the score statistic $S_{\theta_0}(Y)$ in finite samples is not always easy to determine. It is, however, possible to obtain an asymptotic approximation to the critical region of the score test.

Property 15.6: *Under the regularity conditions ensuring the asymptotic normal distribution of the score vector (see Chapter 7), the score test is asymptotically a nonrandomized test with critical region*

$$W = \left\{ S_{\theta_0}(Y) > \sqrt{n\hat{\mathcal{I}}(\theta_0)}z_{1-\alpha} \right\},$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of the standard normal distribution $N(0,1)$ and $\hat{\mathcal{I}}(\theta_0)$ is a consistent estimator of the Fisher information $\mathcal{I}(\theta_0)$ corresponding to one observation.

PROOF: This property follows from the form of the score test and the fact that

$$\frac{S_{\theta_0}(Y)}{\sqrt{n\hat{\mathcal{I}}(\theta_0)}} \xrightarrow{d} N(0,1),$$

as $n \to \infty$. □

Property 15.6 is frequently easy to use. As illustrations, we consider a sampling model and a conditional model.

a) In the case of a sampling model, we have

$$\ell(y;\theta) = \prod_{i=1}^{n} f(y_i;\theta),$$

where f is the density of one observation. The score $S_\theta(Y)$ is equal to

$$S_\theta(Y) = \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial\theta}.$$

The Fisher information for one observation (say Y_1) is

$$\mathcal{I}(\theta_0) = -E_{\theta_0}\left(\frac{\partial^2 \log f(Y_1;\theta_0)}{\partial\theta^2}\right) = E_{\theta_0}\left(\frac{\partial \log f(Y_1;\theta_0)}{\partial\theta}\right)^2.$$

Such a quantity is known because θ_0 is known. Thus we can let $\hat{\mathcal{I}}(\theta_0) = \mathcal{I}(\theta_0)$.

In this case, the score test corresponds to the critical region

$$W = \left\{ \sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial \theta} \geq \sqrt{nE_{\theta_0}\left(\frac{\partial \log f(Y_1;\theta_0)}{\partial \theta}\right)^2} z_{1-\alpha} \right\}.$$

Other score tests can be obtained by replacing the expectation in the Fisher information by the sample mean in the latter expression. Thus another score test corresponds to the critical region

$$\tilde{W} = \left\{ \sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial \theta} \geq \sqrt{\sum_{i=1}^{n}\left(\frac{\partial \log f(y_i;\theta_0)}{\partial \theta}\right)^2} z_{1-\alpha} \right\}.$$

The latter procedure reduces to comparing the usual mean of the individual scores to the root mean of the squared individual scores.

b) In the case of a conditional model, we have

$$\ell(y;\theta) = \prod_{i=1}^{n} f(y_i \mid x_i;\theta), \qquad S_{\theta_0}(y) = \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid x_i;\theta_0)}{\partial \theta},$$

and

$$\mathcal{I}_c(\theta_0) = -E_{\theta_0}\left(\frac{\partial^2 \log f(Y \mid X;\theta_0)}{\partial \theta^2}\right) = E_{\theta_0}\left(\frac{\partial \log f(Y \mid X;\theta_0)}{\partial \theta}\right)^2.$$

Here, the Fisher information is typically unknown since the probability distribution of X is unknown. However, it can be estimated consistently by

$$\hat{\mathcal{I}}(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \log f(Y_i \mid X_i;\theta_0)}{\partial \theta^2}\right),$$

or by

$$\tilde{\mathcal{I}}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \log f(Y_i \mid X_i;\theta_0)}{\partial \theta}\right)^2,$$

To each of these consistent estimators corresponds a critical region for the score test. Namely, we have

$$\hat{W} = \left\{ \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid x_i;\theta)}{\partial \theta} \geq \sqrt{-\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i \mid x_i;\theta_0)}{\partial \theta^2}} z_{1-\alpha} \right\}.$$

and

$$\tilde{W} = \left\{ \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid x_i; \theta_0)}{\partial \theta} \geq \sqrt{\sum_{i=1}^{n} \left(\frac{\partial \log f(y_i \mid x_i; \theta_0)}{\partial \theta} \right)^2} z_{1-\alpha} \right\}.$$

Example 15.4: Consider a random sample of size n drawn from the Cauchy distribution with density

$$f(y_i; \theta) = \frac{1}{\pi} \frac{1}{(1 + (y_i - \theta)^2)}, \ \theta \in I\!R, \ i = 1, \ldots, n.$$

This distribution is symmetric about θ. Thus θ can be interpreted as the median (recall that the mean does not exist). The conditions of Property 15.3 are not satisfied. Property 15.5 allows us, however, to construct a LUMP test at level α for testing H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$.

The score test is given by

$$S_{\theta_0}(Y) = 2 \sum_{i=1}^{n} \frac{Y_i - \theta_0}{1 + (Y_i - \theta_0)^2},$$

which has zero mean, i.e., $E_{\theta_0} S_{\theta_0}(Y) = 0$. Thus Fisher information is

$$\mathcal{I}(\theta_0) = E_{\theta_0} \left(\frac{2(Y_i - \theta_0)}{1 + (Y_i - \theta_0)^2} \right)^2 = \frac{1}{2}.$$

Therefore

$$\sqrt{\frac{2}{n}} S_{\theta_0}(Y) \xrightarrow{d} N(0, 1).$$

Hence, asymptotically, the test with critical region

$$W = \left\{ 2 \sum_{i=1}^{n} \frac{(y_i - \theta_0)}{1 + (y_i - \theta_0)^2} \geq \sqrt{\frac{n}{2}} z_{1-\alpha} \right\}$$

is a LUMP test.

15.3 Tests of Two-Sided Hypotheses

We consider again the one-parameter model $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$, where Θ is an interval of $I\!R$. We consider now testing situations, where either the null hypothesis or the alternative hypothesis is two-sided. Specifically,

we consider problems defined by null hypotheses of the form H_0: $\theta \in \Theta_0$, where

$$\Theta_0 = \{\theta : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2\},$$

or

$$\Theta_0 = \{\theta : \theta_1 \leq \theta \leq \theta_2\},$$

or

$$\Theta_0 = \{\theta_0\}.$$

The alternative hypotheses are defined by H_1: $\{\theta \in \Theta_1 = {}^c\Theta_0\}$.

We distinguish two cases.

15.3.1 Case Where $\Theta_0 = \{\theta : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2\}$

In fact, this type of null hypotheses is seldom considered in practice. In this case, however, there exists a UMP test at level α when the model is exponential. We give below the form of the UMP test without proof (see Lehmann (1959, p.88) for detail).

Property 15.7: *Assume that* $\ell(y;\theta) = C(\theta)h(y)\exp(Q(\theta)T(y))$, *where* Q *is strictly increasing. Consider the testing problem defined by* $\Theta_0 = \{\theta : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2\}$ *and* $\Theta_1 = {}^c\Theta_0$. *Then there exists a UMP test at level* α, $0 < \alpha < 1$. *It is given by*

$$\begin{cases} \phi(y) = 1, & \text{if } c_1 < T(y) < c_2, \\ \phi(y) = 0, & \text{if } T(y) < c_1 \text{ or } T(y) > c_2, \\ \phi(y) = \lambda_i, & \text{if } T(y) = c_i,\ i = 1, 2, \end{cases}$$

where c_i *and* γ_i, $i = 1, 2$, *are determined by the two conditions*

$$E_{\theta_1}\phi = E_{\theta_2}\phi = \alpha.$$

15.3.2 Case Where $\Theta_0 = \{\theta : \theta_1 \leq \theta \leq \theta_2\}$

When Θ_0 is an interval, possibly reduced to a point, there does not exist a UMP test at level α.

To see this, consider the favorable case where the model is exponential with $Q(\theta)$ strictly increasing. If there exists a UMP test ϕ at level α for testing Θ_0 against $\Theta_1 = {}^c\Theta_0$, then this test would also be UMP at level α for testing (i) Θ_0 against $\Theta_1^* = \{\theta : \theta > \theta_2\}$, and (ii) Θ_0 against

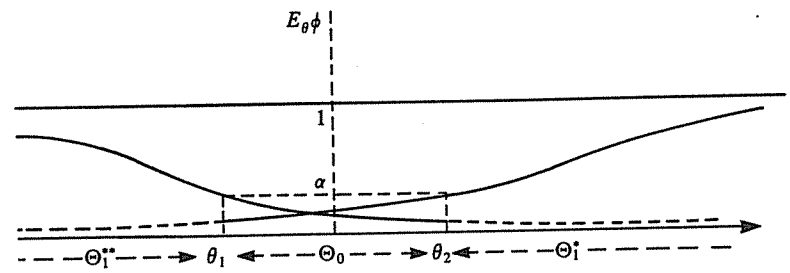


Figure 15.5: Graph of $E_\theta(\phi)$

$\Theta_1^{**} = \{\theta : \theta < \theta_1\}$. This would imply that $E_\theta\phi$ is simultaneously strictly increasing and strictly decreasing on Θ_0 (see Exercise 15.5), which is impossible. See Figure 15.5.

Remark 15.6: Note, however, that the power of a UMP test at level α for testing Θ_0 against $\Theta_1^* = \{\theta : \theta > \theta_2\}$, say, is smaller than α when $\theta < \theta_1$. Therefore, when the alternative hypothesis is $\Theta_1 = \Theta_1^* \cup \Theta_1^{**}$, where $\Theta_1^{**} = \{\theta : \theta < \theta_1\}$, the probability of rejecting H_0 when H_0 is false can be smaller than the probability of rejecting H_0 when H_0 is true. Hence it is natural to exclude this type of tests. In the next chapter we shall study this type of exclusion from the set of possible tests.

15.4 Exercises

EXERCISE 15.1: Show that the intersection of the risk diagram considered in Section 15.1.1 with the boundary of the unit square $[0,1] \times [0,1]$ is reduced to the two points $(0,1)$ and $(1,0)$ if and only if $\{A : P_0(A) = 0\} = \{A : P_1(A) = 0\}$, i.e., if and only if the probability distributions P_0 and P_1 are *equivalent.*

EXERCISE 15.2: Let ϕ be an optimal test at level α, $0 < \alpha < 1$, for testing a simple null hypothesis against a simple alternative hypothesis. Show that $E_1\phi = \alpha$ if and only if $P_1 = P_0$.

EXERCISE 15.3: Consider the sampling model

$$\left(\mathbb{R}^n, \left\{\mathcal{U}_{[0,\theta]}^{\otimes n},\ \theta = 2 \text{ or } 3\right\}\right),$$

where $\mathcal{U}_{[0,\theta]}$ is the uniform distribution on $[0,\theta]$. We want to test H_0: $\theta = 2$ against H_1: $\theta = 3$.

a) Show that there exists a sufficient statistic for θ.

b) Find an optimal test at level α in the Neyman sense. Is this test unique?

c) Compute the power of this test.

d) Draw the risk diagram.

EXERCISE 15.4: Consider Example 15.2.

a) Find the equation of the lower boundary of the risk diagram.

b) Let ϕ be an optimal test in the Neyman sense at level α. Find the prior distribution for which this test is also optimal in the Bayesian sense.

c) Study how the risk diagram changes when $n \to \infty$, when $\theta_1 \to +\infty$, and when $\theta_1 \to \theta_0$.

EXERCISE 15.5: Let $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$ be an identified model, where Θ is an interval of $\mathbb{R}$. Suppose that the family of probability distributions $\{P_\theta, \theta \in \Theta\}$ is a monotone likelihood ratio family in a statistic U. Let ϕ be a test defined by $\phi = 1$ if $U > c$, $\phi = 0$ if $U < c$, and $\phi = \gamma$ if $U = c$. Using the result of Exercise 15.2, show that the function $E_\theta \phi$ is strictly increasing in θ.

EXERCISE 15.6: Consider the model defined by

$$\mathcal{Y} = [0, 2], \quad P_0 = f_0 \cdot \lambda, \quad P_1 = f_1 \cdot \lambda,$$

where

$$f_0(y) = \begin{cases} y, & \text{if } 0 \leq y \leq 1, \\ 2 - y, & \text{if } 1 \leq y \leq 2, \end{cases}$$

and

$$f_1(y) = 1/2.$$

One wants to test P_0 against P_1.

a) Consider critical regions W of the form $[0, y)$. Determine $\alpha(W)$ and $\beta(W)$ as functions of y. Draw the risk diagram associated with the critical regions W.

b) Find an optimal test at level α. Compute its power when $\alpha = 0.05$.

c) In the space (α, β), represent the family of most powerful tests when α varies. Determine the set of admissible tests and the set of *minimax tests*. The minimax tests are the tests that minimize the maximum risk.

EXERCISE 15.7: Consider the model $(\mathcal{Y}, \{P_\theta, \Theta \in [0,1]\})$, where $\mathcal{Y} = [0,2]$ and P_θ is the probability distribution whose density on $\mathcal{Y}$ is

$$\begin{cases} (2\theta - 1)y + (1 - \theta), & \text{if } y \in [0,1], \\ (1 - 2\theta)y + (3\theta - 1), & \text{if } y \in [1,2]. \end{cases}$$

Show that this family of densities has the monotone likelihood ratio property. Then find a UMP test for testing H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$.

EXERCISE 15.8: Consider two probability distributions P_0 and P_1 with disjoint supports, i.e., such that $\exists A \subset \mathcal{Y} : P_0(A) = 1$ and $P_1({}^cA) = 1$. Discuss the problem of testing the null hypothesis H_0: $P = P_0$ against the alternative hypothesis H_1: $P = P_1$. Does the form of the Neyman test appear to agree with the intuition?

15.5 References

Ferguson, T.S. (1967). *Mathematical Statistics*, Academic Press.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, Wiley.

Monfort, A. (1982). *Cours de Statistique Mathématique*, Economica.

CHAPTER 16

Unbiased Tests and Invariant Tests

When the parametric model has more than one parameter, there does not exist, in general, a uniformly most powerful test or even a locally uniformly most powerful test. Then "optimal" tests must be found in some restricted classes of test functions. In this chapter we consider two such classes. These are the class of unbiased tests and the class of invariant tests.

16.1 Unbiased Tests

16.1.1 Unbiased Tests and α-Similar Tests

We consider the parametric model $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$ and the testing problem defined by H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta_1 = {}^c\Theta_0$. We assume that Θ is an open subset of $\mathbb{R}^p$.

Definition 16.1: *A test ϕ is unbiased at level α if*

(i) $E_\theta\phi \leq \alpha$, $\forall\theta \in \Theta_0$, *i.e., the level of the test ϕ is less or equal to α and*

(ii) $E_\theta\phi \geq \alpha$, $\forall\theta \in \Theta_1$, *i.e., the power of the test ϕ is larger or equal to α under the alternative hypothesis.*

Thus, if the test is unbiased, the probability of accepting H_1 is at least as large when H_1 is true as H_0 is true. When the parameter θ is a scalar and when the hypotheses are one-sided, we noted in Chapter 15 that it was natural to restrict ourselves to tests that are unbiased so as to find an optimal test in the Neyman sense in this restricted class of tests.

Definition 16.2: *A test ϕ is uniformly most powerful and unbiased (UMPU) at level α if*

(i) it is unbiased at level α and

(ii) it is more powerful than any other unbiased test of level α.

Note that condition (i) in Definition 16.2 can be replaced by the condition that ϕ is of level less than or equal to α. That is, the unbiasedness requirement is unnecessary. This is because the constant test $\phi' = \alpha$ is unbiased. Thus, condition (ii) implies $E_\theta\phi \geq E_\theta\phi' = \alpha$, $\forall\ \theta \in \Theta_1$, i.e., condition (i).

Remark 16.1: Similarly, it is obvious that a UMP test at level α is necessarily unbiased. This is because the power of such a test is greater than the constant test $\phi' = \alpha$. Therefore any UMP test at level α is UMPU at this level α.

It is useful to define another class of tests, namely the class of α-similar tests.

Definition 16.3: *A test ϕ is α-similar on a subset Θ^* of Θ if $E_\theta\phi = \alpha$, $\forall\ \theta \in \Theta^*$.*

In this chapter, we are interested in tests that are α-similar on the common boundary of the two hypotheses, i.e., on $\Theta_F = \bar{\Theta}_0 \cap \bar{\Theta}_1$, where $\bar{\Theta}_0$ and $\bar{\Theta}_1$ are the closures of Θ_0 and Θ_1 respectively. Such tests arise naturally when considering unbiased tests at level α.

Property 16.1:

(i) Let ϕ^* be an unbiased test at level α. Suppose that the function $E_\theta\phi^*$ is continuous at any point of Θ_F. Then ϕ^* is an α-similar test on Θ_F.

(ii) Suppose that the function $E_\theta\phi$ is continuous at any point of Θ_F for every test ϕ. Then a test ϕ^* that is uniformly most powerful in the class of α-similar tests on Θ_F is necessarily uniformly most powerful and unbiased at level α.

PROOF:

(i) Every point θ in Θ_F is the limit of a sequence $\{\theta_{0n}\}$ of points in Θ_0 and the limit of a sequence $\{\theta_{1n}\}$ of points in Θ_1. The continuity in θ of the mapping $E_\theta\phi^*$ implies

$$E_\theta\phi^* = \lim_{n\to\infty} E_{\theta_{0n}}\phi^* \leq \alpha$$

and

$$E_\theta \phi^* = \lim_{n \to \infty} E_{\theta_{1n}} \phi^* \geq \alpha.$$

Thus $E_\theta \phi^* = \alpha$, $\forall\ \theta \in \Theta_F$. Therefore ϕ^* is an α-similar test on Θ_F.

(ii) The second part of Property 16.1 is a direct consequence of the first part, which implies that the class of α-similar tests contains the class of unbiased tests at level α. □

The continuity of $E_\theta \phi^*$ mentioned in Property 16.1 is not satisfied in all models . In particular, it is satisfied in exponential models. These models are considered in the next subsection.

16.1.2 UMPU Tests in the Absence of Nuisance Parameters

First, we consider a one-parameter exponential model $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$, where Θ is an open interval of $I\!R$. The densities are

$$\ell(y; \theta) = C(\theta) h(y) \exp(Q(\theta) T(y)). \tag{16.1}$$

The function $Q(\theta)$ is assumed to be strictly increasing and continuous in θ.

We consider the problem of testing the null hypothesis defined by $\Theta_0 = \{\theta : \theta_1 \leq \theta \leq \theta_2\}$ against the two-sided alternative hypothesis defined by $\Theta_1 = {}^c\Theta_0$, where $\theta_1, \theta_2 \in \Theta$. For this type of testing problems, we know that there does not exist a UMP test (see Section 15.3.2). On the other hand, there exists a UMPU test at level α, as the next property states.

Property 16.2: *Within the preceding framework, there exists a UMPU test at level α, $0 < \alpha < 1$, for testing the null hypothesis H_0: $\theta_1 \leq \theta \leq \theta_2$ against the two-sided alternative hypothesis H_1: $\theta \in {}^c\Theta_0$. This test is defined by*

$$\begin{cases} \phi(y) = 1, & \text{if } T(y) < c_1 \text{ or } T(y) > c_2, \\ \phi(y) = 0, & c_1 < T(y) < c_2, \\ \phi(y) = \gamma_i, & \text{if } T(y) = c_i, \quad i = 1, 2, \end{cases}$$

where the constants c_i, and γ_i, $i = 1, 2$, are determined by the two conditions

$$E_{\theta_1} \phi = E_{\theta_2} \phi = \alpha.$$

SKETCH OF THE PROOF: The proof consists of three steps.

a) First, we use Property 15.7. From this property it follows that there exists a uniformly most powerful test ψ (say) at level $1-\alpha$ for testing

$$\tilde{H}_0 : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2$$

against

$$\tilde{H}_1 : \theta_1 < \theta < \theta_2.$$

From Property 15.7 it also follows that the form of the test ψ is identical to that of the test $\psi = 1 - \phi$, where ϕ is the test given in Property 16.2.

b) Because it is optimal, the test ψ must be preferred to the constant test equal to $1-\alpha$. Thus we have

$$E_\theta \psi = E_\theta(1-\phi) \geq 1-\alpha, \ \forall\, \theta, \ \theta_1 < \theta < \theta_2.$$

It follows that $E_\theta \phi \leq \alpha$, $\forall\, \theta$, $\theta_1 < \theta < \theta_2$. Hence the level of ϕ is less or equal to α.

c) To complete the proof, it suffices to verify that ϕ is optimal in the class of α-similar tests on $\Theta_F = \{\theta_1, \theta_2\}$, i.e., to show that the test ψ minimizes $E_\theta \psi^*$ subject to the constraints $E_{\theta_1}\psi^* = E_{\theta_2}\psi^* = 1-\alpha$ for every value $\theta \in \Theta_1$. □

Property 16.2 can be generalized to the limit case where $\theta_1 = \theta_2$.

Corollary 16.1: *Suppose that the family of densities is of the form*

$$\ell(y;\theta) = C(\theta)h(y)\exp(Q(\theta)T(y)),$$

where Q is continuous and strictly increasing in θ. Then there exists a UMPU test for testing H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$, where θ_0 is an interior point of Θ. This test is defined by

$$\begin{cases} \phi(y) = 1, & \text{if } T(y) < c_1 \text{ or } T(y) > c_2, \\ \phi(y) = 0, & c_1 < T(y) < c_2, \\ \phi(y) = \gamma_i, & \text{if } T(y) = c_i, \quad i = 1,2, \end{cases}$$

where the constants c_i and γ_i are determined by the conditions

$$E_{\theta_0}\phi = \alpha \text{ and } E_{\theta_0}(\phi T) = \alpha E_{\theta_0} T.$$

SKETCH OF THE PROOF: First, we use Property 16.2, where we let θ_2 converge to θ_1. Given that the form of the optimal test remains the

same, we basically need to study how the conditions determining the constants c_i and γ_i, $i = 1, 2$, change when $\theta_2 \to \theta_1$.

Now the conditions of Property 16.2 can be written as

$$\begin{cases} E_{\theta_1}\phi = \alpha, \\ E_{\theta_2}\phi - E_{\theta_1}\phi = 0, \end{cases}$$

i.e.

$$\begin{cases} E_{\theta_1}\phi = \alpha, \\ (E_{\theta_2}\phi - E_{\theta_1}\phi)/(\theta_2 - \theta_1) = 0. \end{cases}$$

Let $\theta_1 = \theta_0$ and $\theta_2 = \theta_0 + \varepsilon$, where ε converges to zero. Then, at the limit, the preceding constraints become

$$E_{\theta_0}\phi = \alpha \text{ and } \left(\frac{\partial E_\theta \phi}{\partial \theta}\right)_{\theta=\theta_0} = 0.$$

Next, it remains to evaluate $\partial E_\theta \phi / \partial \theta$. We have

$$\begin{aligned} \frac{\partial E_\theta \phi}{\partial \theta} &= \frac{\partial C(\theta)}{\partial \theta} \int_{\mathcal{Y}} \phi(y) h(y) \exp(Q(\theta)T(y)) d\mu(y) \\ &\quad + C(\theta) \frac{\partial Q(\theta)}{\partial \theta} \int_{\mathcal{Y}} \phi(y) h(y) T(y) \exp(Q(\theta)T(y)) d\mu(y) \\ &= \frac{\partial \log C(\theta)}{\partial \theta} E_\theta \phi + \frac{\partial Q(\theta)}{\partial \theta} E_\theta(\phi T). \end{aligned}$$

But, because the score vector has zero mean, we also have

$$E_\theta \frac{\partial \log \ell(Y;\theta)}{\partial \theta} = 0 = \frac{\partial \log C(\theta)}{\partial \theta} + \frac{\partial Q(\theta)}{\partial \theta} E_\theta(T).$$

It follows that the equality

$$\left(\frac{\partial E_\theta \phi}{\partial \theta}\right)_{\theta=\theta_0} = 0$$

can be written as

$$E_{\theta_0}(\phi T) = E_{\theta_0}\phi \; E_{\theta_0} T = \alpha E_{\theta_0} T.$$

□

Remark 16.2: The two conditions determining the constants c_i and γ_i in Corollary 16.1 imply the orthogonality between the test ϕ and the sufficient statistic T under H_0 in the sense that

$$\mathrm{Cov}_{\theta_0}(\phi, T) = E_{\theta_0}(\phi T) - E_{\theta_0}\phi E_{\theta_0} T = 0.$$

Remark 16.3: In many cases, the probability distribution of the statistic $T(Y)$ under H_0: $\theta = \theta_0$ is symmetric about a real number a. In those cases, it is possible to obtain a simpler form for the optimal test in the class of unbiased tests at level α. Namely, the test becomes

$$\begin{cases} \phi(y) = 1, & \text{if } |T(y) - a| > c, \\ \phi(y) = 0, & \text{if } |T(y) - a| < c, \\ \phi(y) = \gamma, & \text{if } |T(y) - a| = c, \end{cases}$$

where c and γ are determined by $E_{\theta_0}\phi = \alpha$.

To see this, note that this test is clearly of the form given in Corollary 16.1. Thus it suffices to verify whether the condition $E_{\theta_0}(\phi T) = E_{\theta_0}\phi E_{\theta_0}T$ is satisfied. This equality follows from

$$E_{\theta_0}(\phi T) = E_{\theta_0}(\phi(T-a)) + aE_{\theta_0}\phi = E_{\theta_0}(\phi(T-a)) + E_{\theta_0}TE_{\theta_0}\phi.$$

But $E_{\theta_0}(\phi(T-a)) = 0$ because $\phi(T-a)$ is an odd function of $T-a$, whose probability distribution is symmetric about zero.

Example 16.1: Consider a random sample $Y_1, \ldots, Y_n$ drawn from the normal distribution $N(\theta, 1)$ with mean $\theta \in \mathbb{R}$. We want to test the hypothesis that the mean is equal to a given value θ_0. Specifically, we want to test H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$.

The model is exponential with densities

$$\ell(y;\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{n\theta^2}{2} - \frac{1}{2}\sum_{i=1}^{n} y_1^2\right) \exp(n\theta\bar{y}).$$

The function $Q(\theta) = \theta$ is continuous and strictly increasing. Thus there exists a UMPU test based on the statistic $T(Y) = \bar{Y}$. Moreover, under the null hypothesis H_0, the probability distribution of $\bar{Y}$ is the normal distribution $N(\theta_0, 1/n)$, which is symmetric about θ_0. Hence a UMPU test is the nonrandomized test with critical region

$$W = \{|\bar{Y} - \theta_0| > c\},$$

where the constant c is determined by the condition

$$P_{\theta_0}(W) = \alpha,$$

i.e.

$$P_{\theta_0}(|\bar{Y} - \theta_0| > c) = \alpha,$$

i.e.

$$P_{\theta_0}(\sqrt{n}|\bar{Y} - \theta_0| > \sqrt{n}c) = \alpha.$$

Since $\sqrt{n}(\bar{Y} - \theta_0)$ follows a standard normal distribution then $\sqrt{n}c$ is equal to the $(1 - \alpha/2)$-th quantile of this distribution. Let this quantile be denoted $z_{1-\alpha/2}$. Thus the critical region of the UMPU test becomes

$$W = \{|\bar{Y} - \theta_0| > \frac{1}{\sqrt{n}} z_{1-\alpha/2}\}.$$

That is, the null hypothesis is rejected if the usual estimate $\bar{Y}$ of the mean θ is sufficiently far from θ_0.

In the special case where $\alpha = 5\%$, we have $z_{1-\alpha/2} = 1.96 \approx 2$. Thus the critical region of the UMPU test is approximately given by

$$W \approx \{|\bar{Y} - \theta_0| > \frac{2}{\sqrt{n}}\}.$$

16.1.3 UMPU Tests in the Presence of Nuisance Parameters

Frequently, one is interested in testing one parameter while this parameter is not the only parameter of the model. In other words, the problem is that of testing a one-parameter hypothesis in a multidimensional parameter model.

Example 16.2: To study the consumption behavior of a group of households, one considers the following model

$$\log C_t = a + b \log R_t + u_t, \quad t = 1, \ldots, T,$$

where C_t is the aggregated household consumption level at time t, R_t is household income, and u_t is a random error. It is assumed that the errors u_t are independently and identically distributed as $N(0, \sigma^2)$.

The income values R_t are considered nonrandom. Alternatively, the model can be viewed as specifying the conditional distribution of $\log C_t$, $t = 1, \ldots, T$, given $R_t, t = 1, \ldots, T$. This conditional distribution is $\prod_{t=1}^{T} N(a + b \log R_t, \sigma^2)$.

The parameters a, b, and σ^2 are unknown. The hypothesis of interest is $H_0 : b \leq 1$. Thus one wants to know whether the income elasticity of consumption is smaller or larger than one, i.e., whether the commodity is an inferior good (hypothesis H_0) or a superior good (hypothesis H_1).

The test procedure discussed up to now no longer applies since b is not the only parameter of the model.

To study such kinds of testing problems, we need another definition and a preliminary result.

Definition 16.4: *Let $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$ be a parametric model and S a sufficient statistic in the submodel $(\mathcal{Y}, \{P_\theta, \theta \in \Theta^*\})$, where $\Theta^* \subset \Theta$. A test ϕ is said to have a Neyman α-structure relative to S and Θ^* if*

$$E_\theta(\phi \mid S) = \alpha, \quad \forall\, \theta \in \Theta^*.$$

Property 16.3:

(i) A test ϕ that has a Neyman α-structure relative to S and Θ^* is α-similar on Θ^*.

(ii) Furthermore, if S is complete in the submodel $(Y, \{P_\theta, \theta \in \Theta^*\})$, then every α-similar test on Θ^* has a Neyman α-structure relative to S and Θ^*.

PROOF:

(i) Let ϕ be a test with a Neyman α-structure relative to S and Θ^*. We have $E_\theta(\phi \mid S) = \alpha$, $\forall \theta \in \Theta^*$, and, by integration, $E_\theta \phi = E_\theta E_\theta(\phi \mid S) = \alpha$, $\forall \theta \in \Theta^*$. Thus the test ϕ is α-similar on Θ_*.

(ii) Conversely, if ϕ is an α-similar test on Θ^*, we have

$$E_\theta \phi = E_\theta E(\phi \mid S) = \alpha, \quad \forall \theta \in \Theta^*.$$

Thus

$$E_\theta \left(E(\phi \mid S) - \alpha\right) = 0, \quad \forall \theta \in \Theta^*.$$

From the definition of completeness, it follows that $E(\phi \mid S) = \alpha$, $\forall \theta \in \Theta_*$, where the subscript θ can be omitted from the second conditional expectation because S is sufficient in the submodel. □

As seen below, the preceding definition and property are directly applicable to exponential models with many parameters.

Hereafter, it is assumed that the chosen parameterization of the multiparameter exponential model corresponds to its canonical parameterization so that its densities can be written as

$$\ell(y; \theta) = C(\theta) \exp\left(\theta^* T(y) + \sum_{i=1}^{p-1} \lambda_i S_i(y)\right), \tag{16.2}$$

with $\theta = (\theta^*, \lambda_1, \ldots, \lambda_{p-1})'$. We let $\lambda = (\lambda_1, \ldots, \lambda_{p-1})'$ and $S = (S_1, \ldots, S_{p-1})'$. The parameter θ^* is the scalar parameter of interest on which the hypothesis is bearing, and λ is the vector of nuisance parameters.

The statistic $(T, S')'$ is sufficient for θ (see Section 3.1.2, Volume 1). Thus we can consider the model induced by this statistic. The probability distributions of the induced model have densities of the form

$$\ell(t, s; \theta) = \tilde{C}(\theta) \exp\left(\theta^* t + \sum_{i=1}^{p-1} \lambda_i s_i\right).$$

It is then easy to verify that the conditional distribution of T given $S = s$ has a density of the form

$$\ell(t \mid s; \theta) = C_s(\theta^*) \exp(\theta^* t)$$

(see Exercise 16.1).

The important point is that this conditional distribution does not depend on the nuisance parameter λ. Therefore we can apply the results on one-parameter models, which are established in Section 16.1.2, to the conditional model for T given S.

a) Tests of One-Sided Hypotheses

First, we consider the problem of testing a one-sided hypothesis about the scalar parameter of interest θ^*. Specifically, we consider testing

$$H_0 : \theta^* \leq \theta_0^* \quad \text{against} \quad H_1 : \theta^* > \theta_0^*.$$

It is assumed that Θ is an open subset and that θ_0^* is a possible value for θ^*.

Property 16.4: *Consider the exponential model (16.2). Then there exists a UMPU test at level α for testing the hypothesis H_0: $\theta^* \leq \theta_0^*$ against the hypothesis H_1: $\theta^* > \theta_0^*$. The test is given by*

$$\begin{cases} \phi(t, s) = 1, & \text{if } t > c(s), \\ \phi(t, s) = 0, & \text{if } t < c(s), \\ \phi(t, s) = \gamma(s), & \text{if } t = c(s), \end{cases}$$

where the functions $c(s)$ and $\gamma(s)$ are determined by the conditions

$$E_{\theta_0^*}(\phi(T, S) \mid S = s) = \alpha, \quad \forall s.$$

PROOF: The proof consists of two steps.

a) THE TEST ϕ IS UNBIASED AT LEVEL α:

To show this, we consider the conditional model for T given $S = s$, where s is arbitrary. This model is parameterized by a single parameter θ^*. Thus, from Property 15.3 and Section 15.2.2, it follows that the test $\phi(T, s)$ is UMP at level α for testing H_0: $\theta^* \leq \theta_0^*$ against H_1: $\theta^* > \theta_0^*$. Moreover, from Remark 16.1 for every given s, the test $\phi(T, s)$ is conditionally unbiased at level α, i.e.

$$\begin{cases} E_\theta(\phi(T,s) \mid S = s) \leq \alpha, & \forall \theta \in \Theta_0 = \{\theta : \theta^* \leq \theta_0^*\}, \\ E_\theta(\phi(T,s) \mid S = s) \geq \alpha, & \forall \theta \in \Theta_1 = \{\theta : \theta^* > \theta_0^*\}. \end{cases}$$

Integrating, we obtain

$$\begin{cases} E_\theta \phi(T,S) \leq \alpha, & \forall \theta \in \Theta_0, \\ E_\theta \phi(T,S) \geq \alpha, & \forall \theta \in \Theta_1, \end{cases}$$

i.e., the test $\phi(T, S)$ is unconditionnally unbiased at level α, as required.

b) THE TEST ϕ IS OPTIMAL:

To show this, we use the fact that the statistic S is complete and sufficient for λ in the exponential model (16.2) when $\theta \in \Theta_0^* = \{\theta : \theta^* = \theta_0^*\}$. Applying Property 16.1, which says that any unbiased test is α-similar, and Property 16.3, which says that any α-similar test has a Neyman α-structure, it follows that a test that is more powerful than any tests with a Neyman α-structure relative to S and Θ_0^* must be UMPU.

Now let $\psi(T, S)$ be a test with a Neyman α-structure relative to S and Θ_0^*, i.e., such that $E_{\theta_0^*}(\psi(T,S) \mid S = s) = \alpha$, $\forall s$. The power of such a test is

$$E_\theta \psi = \int \left(\int \psi(t,s) dP_{\theta^*}^{T|S=s}(t) \right) dP_\theta^S(s), \quad \theta \in \Theta_1.$$

Since, for every given s, the test ϕ is UMP at level α in the conditional model (see a)), then ϕ maximizes the quantity in parentheses in the preceding expression for every s. Integrating, we obtain

$$E_\theta \phi \geq E_\theta \psi, \quad \forall \theta \in \Theta_1.$$

□

It is not always easy to find functions $c(s)$ and $\gamma(s)$ corresponding to a UMPU test. The determination of such functions may become easier

by invoking the next result, which replaces the functions $c(s)$ and $\gamma(s)$ by two constants c and γ.

Property 16.5: *Under the assumptions of Property 16.4, if $U = f(T,S)$ is an ancillary statistic for $\theta \in \Theta_0^* = \{\theta : \theta^* = \theta_0^*\}$ and if f is strictly increasing in t for every given s, then the test defined by*

$$\begin{cases} \tilde{\phi}(u) = 1, & \text{if } u > c, \\ \tilde{\phi}(u) = 0, & \text{if } u < c, \\ \tilde{\phi}(u) = \gamma, & \text{if } u = c, \end{cases}$$

with $E_{\theta_0^*}\tilde{\phi} = \alpha$, where $u = f(t,s)$, is UMPU at level α for testing the hypothesis

$$H_0 : \theta^* \leq \theta_0^* \quad \text{against} \quad H_1 : \theta^* > \theta_0^*.$$

PROOF: The proof consists of three steps.

a) The proof of the existence of the test $\tilde{\phi}$ is similar to the proof of the Neyman–Pearson Theorem.

b) Given that the function f is increasing in t, the test $\tilde{\phi}$ can be written as

$$\begin{cases} \tilde{\phi}(f(t,s)) = 1, & \text{if } t > c(s), \\ \tilde{\phi}(f(t,s)) = 0, & \text{if } t < c(s), \\ \tilde{\phi}(f(t,s)) = \gamma, & \text{if } t = c(s). \end{cases}$$

This form corresponds to an optimal test (see Property 16.4).

c) It remains to verify that the conditions of Property 16.4 determining $c(s)$ and $\gamma(s)$ hold, i.e., that $\tilde{\phi}$ has a Neyman α-structure relative to S and Θ_0^*. But we know that the statistic S is complete and sufficient for $\theta \in \Theta_0^*$ and that $U = f(T,S)$ is ancillary for $\theta \in \Theta_0^*$. From Property 3.6, it follows that these two statistics are independent when $\theta \in \Theta_0^*$. This implies that $\alpha = E_{\theta_0^*}\tilde{\phi} = E_{\theta_0^*}(\tilde{\phi} \mid S = s)$ because $\tilde{\phi}$ depends on U only. □

b) Tests of Two-Sided Hypotheses

Similar results can be obtained when the null hypothesis is of the form H_0: $\theta_1^* \leq \theta^* \leq \theta_2^*$ or H_0^*: $\theta^* = \theta_0^*$, i.e., when the alternative hypothesis is two-sided. We give below, without proof, properties that are analogous to those obtained in the one-sided case.

Property 16.6: *Consider the exponential model (16.2). Then there exists a UMPU test at level α for testing the hypothesis $H_0 : \theta_1^* \leq \theta^* \leq \theta_2^*$ against $H_1 : \theta^* < \theta_1^*$ or $\theta^* > \theta_2^*$. The test is given by*

$$\begin{cases} \phi(t,s) = 1, & \text{if } t < c_1(s) \text{ or } t > c_2(s), \\ \phi(t,s) = 0, & \text{if } c_1(s) < t < c_2(s), \\ \phi(t,s) = \gamma_i(s), & \text{if } t = c_i(s) \quad i = 1,2, \end{cases}$$

where the functions $\gamma_i(s)$ and $c_i(s)$ are determined by the conditions

$$E_{\theta_i^*}(\phi(T,S) \mid S = s) = \alpha, \quad \forall s, \ i = 1,2.$$

When the null hypothesis is H_0: $\theta^* = \theta_0^*$, we can proceed as in Corollary 16.1 by letting $\theta_1^* = \theta_0^*$ and $\theta_2^* = \theta_0^* + \varepsilon$ with ε converging to zero.

Property 16.7 *Consider the exponential model (16.2). Then there exists a UMPU test at level α for testing the hypothesis $H_0^* : \theta^* = \theta_0^*$ against $H_1^* : \theta^* \neq \theta_0^*$. The test is given by*

$$\begin{cases} \phi(t,s) = 1, & \text{if } t < c_1(s) \text{ or } t > c_2(s), \\ \phi(t,s) = 0, & \text{if } c_1(s) < t < c_2(s), \\ \phi(t,s) = \gamma_i(s), & \text{if } t = c_i(s), \quad i = 1,2, \end{cases}$$

where the functions $c_i(s)$ and $\gamma_i(s)$ are determined by the conditions

$$E_{\theta_0^*}(\phi(T,S) \mid S = s) = \alpha, \ \forall \alpha$$

and

$$E_{\theta_0^*}(T\phi(T,S) \mid S = s) = \alpha E_{\theta_0^*}(T \mid S = s), \ \forall s.$$

As in the one-sided case, the functions $c_1(s), c_2(s), \gamma_1(s)$, and $\gamma_2(s)$ may be replaced by constant functions. This follows from the purpose of the next result.

Property 16.8: *Consider the exponential model (16.2) and the problem of testing a null hypothesis of the form*

$$H_0 : \theta_1^* \leq \theta^* \leq \theta_0^* \text{ or } H_0^* : \theta^* = \theta_0^*.$$

In the case of a null hypothesis of the form H_0, it is assumed that there exists a statistic $U = f(T,S)$ that is ancillary on $\Theta_1^* \equiv \{\theta : \theta^* = \theta_1^*\}$ and

$\Theta_2^* = \{\theta : \theta^* = \theta_2^*\}$ *such that* $f(t,s)$ *is strictly increasing in* t *for every given* s. *In the case of a null hypothesis of the form* H_0^*, *it is assumed that the statistic* $U = f(T,S)$ *is ancillary on* $\Theta_0^* \equiv \{\theta : \theta^* = \theta_0^*\}$ *and that* $f(t,s)$ *is linear and strictly increasing in* t *for every given* s.

Then there exists a UMPU test at level α. *The test is given by*

$$\begin{cases} \tilde{\phi}(u) = 1, & \text{if } u < c_1 \text{ or } u > c_2, \\ \tilde{\phi}(u) = 0, & \text{if } c_1 < u < c_2, \\ \tilde{\phi}(u) = \gamma_i, & \text{if } u = c_i,\ i = 1,2, \end{cases}$$

where $u = f(t,s)$, *and* c_i *and* γ_i, $i = 1,2$, *are determined by the conditions*

$$E_{\theta_i^*}\tilde{\phi} = \alpha, \quad i = 1,2 \text{ when testing } H_0$$

and

$$E_{\theta_0^*}\tilde{\phi} = \alpha,\ E_{\theta_0^*}(U\tilde{\phi}) = \alpha E_{\theta_0^*}U \quad \text{when testing } H_0^*.$$

PROOF: Given that the function f is increasing, then the test $\tilde{\phi}$ is of the form considered in Properties 16.6 and 16.7, namely

$$\begin{cases} \tilde{\phi}(f(t,s)) = 1, & \text{if } t < c_1(s) \text{ or } t > c_2(s), \\ \tilde{\phi}(f(t,s)) = 0, & \text{if } c_1(s) < t < c_2(s), \\ \tilde{\phi}(f(t,s)) = \gamma_i, & \text{if } t = c_i(s),\ i = 1,2. \end{cases}$$

a) In the case of a null hypothesis of the form H_0, U is ancillary on Θ_1^* and Θ_2^*. Thus U is independent of S on Θ_1^* and Θ_2^*. Hence we have

$$E_{\theta_i^*}\tilde{\phi} = E_{\theta_i^*}(\tilde{\phi} \mid S = s) = \alpha, \quad \forall\ s, i = 1,2,$$

Therefore $\tilde{\phi}$ satisfies the optimality conditions of Property 16.6 as required.

b) In the case of a null hypothesis of the form H_0^*, U is ancillary on Θ_0^* and hence independent of S on Θ_0^*. Thus we have

$$E_{\theta_0^*}\tilde{\phi} = E_{\theta_0^*}(\tilde{\phi} \mid S = s) = \alpha, \quad \forall\ s.$$

Moreover, using $U = a(S) + b(S)T$ and the same independence argument, the condition

$$E_{\theta_0^*}(U\tilde{\phi}) = \alpha E_{\theta_0^*}U$$

can be written as

$$E_{\theta_0^*}(a(S)\tilde{\phi}) + E_{\theta_0^*}(b(S)T\tilde{\phi}) = \alpha E_{\theta_0^*}a(S) + \alpha E_{\theta_0^*}(b(S)T),$$

i.e.

$$\alpha E_{\theta_0^*} a(S) + E_{\theta_0^*}(b(S)T\tilde{\phi}) = \alpha E_{\theta_0^*} a(S) + \alpha E_{\theta_0^*}(b(S)T),$$

i.e.

$$E_{\theta_0^*}\left(b(S)E_{\theta_0^*}((T\tilde{\phi} - \alpha T) \mid S)\right) = 0.$$

Since S is complete, we have

$$b(s)E_{\theta_0^*}\left((T\tilde{\phi} - \alpha T) \mid S = s\right) = 0, \quad \forall s.$$

Moreover, since $b(s)$ is strictly positive everywhere, then

$$E_{\theta_0^*}(T\tilde{\phi} \mid S = s) = \alpha E_{\theta_0^*}(T \mid S = s), \quad \forall s,$$

Therefore ϕ satisfies the optimality conditions of Property 16.7. □

c) Tests of a Scalar Linear Hypothesis About the Canonical Parameters of an Exponential Model

In the preceding subsections we have considered hypotheses bearing directly on one of the canonical parameters of an exponential model. The results we have obtained can be easily generalized to tests of one linear restriction on these canonical parameters.

Property 16.9: *Consider an exponential model whith densities of the form*

$$\ell(y;\theta) = C(\theta)h(y)\exp\left(\sum_{j=1}^{p}\theta_j T_j(y)\right).$$

Then there exist UMPU tests at level α for testing null hypotheses of the form

(i) $H_0 : \sum_{j=1}^{p} a_j\theta_j \leq \theta_0^*$,

(ii) $H_0 : \theta_1^* \leq \sum_{j=1}^{p} a_j\theta_j \leq \theta_2^*$,

(iii) $H_0 : \sum_{j=1}^{p} a_j\theta_j = \theta_0^*$,

where a_j, $j = 1,\ldots,p$, and θ_i^*, $i = 0,1,2$ are given real numbers.

PROOF: To establish the desired result, it suffices to show that $\theta^* = \sum_{j=1}^{p} a_j\theta_j$ can be interpreted as a canonical parameter in the exponential model and to invoke Properties 16.4, 16.6, and 16.7 respectively.

Up to a possible reordering of the parameters, we can assume that $a_1 \neq 0$. Then we can rewrite the density as

$$\ell(y;\theta) = C(\theta)h(y)\exp\left(\theta^*\frac{T_1(y)}{a_1} + \sum_{j=2}^{p}\theta_j\left(T_j(y) - \frac{a_j}{a_1}T_1(y)\right)\right).$$

Making the change in parameters

$$\theta \to \begin{pmatrix} \theta^* & = \sum_{j=1}^{p} a_j\theta_j \\ \lambda_j & = \theta_j, \quad j = 2,\ldots,p \end{pmatrix},$$

we obtain an exponential family of the desired form with

$$T^*(y) = \frac{T_1(y)}{a_1}$$

and

$$S^*_{j-1}(y) = T_j(y) - \frac{a_j}{a_1}T_1(y), \quad j = 2,\ldots,p.$$

□

It is important to note that Property 16.9 holds when only one linear restriction on the canonical parameters is under test.

Example 16.3: Let the observations $Y_1,\ldots,Y_K$ be distributed as $M(n;p_1,\ldots,p_K)$. The corresponding family of densities is

$$\begin{aligned}\ell(y;p) &= \frac{n!}{y_1!\ldots y_K!}\prod_{k=1}^{K}p_k^{y_k} \\ &= \frac{n!}{y_1!\ldots y_K!}\exp\left(\sum_{k=1}^{K}y_k\log p_k\right).\end{aligned}$$

The parameters $p_1,\ldots,p_K$ are not independent from each other since $\sum_{k=1}^{K}p_k = 1$. Eliminating one of them, p_K say, we obtain

$$\ell(y;p) = \frac{n!}{y_1!\ldots y_K!}\exp\left(\left(\sum_{k=1}^{K-1}y_k\log p_k\right) + y_K\log\left(1 - \sum_{k=1}^{K-1}p_k\right)\right).$$

Then, using the relation $n = \sum_{k=1}^{K}y_k$, we obtain

$$\begin{aligned}\ell(y;p) &= \frac{n!}{y_1!\ldots y_K!}\times \\ &\quad \exp\left(n\log\left(1 - \sum_{k=1}^{K-1}p_k\right)\right)\exp\left(\sum_{k=1}^{K-1}y_k\log\left(\frac{p_k}{p_K}\right)\right).\end{aligned}$$

From Property 16.9, it follows that there exist UMPU tests for testing hypotheses about scalar linear forms of the canonical parameters $\log p_k/p_K$, $k = 1, \ldots, K-1$. Such a linear form can be written as

$$\sum_{k=1}^{K-1} a_k \log\left(\frac{p_k}{p_K}\right) = \sum_{k=1}^{K-1} a_k \log p_k - \left(\sum_{k=1}^{K-1} a_k\right) \log p_K,$$

or equivalently

$$\sum_{k=1}^{K} a_k \log p_k \quad \text{with} \quad \sum_{k=1}^{K} a_k = 0.$$

In Section 16.1.3-e, we shall derive a test for independence based on such a linear form.

Example 16.4: Consider a random sample $Y_1, \ldots, Y_n$ drawn from the normal distribution $N(m, \sigma^2)$. The densities of the model are

$$\begin{aligned} \ell(y; m, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - m)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{nm^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{nm}{\sigma^2}\bar{y}\right). \end{aligned}$$

This family of densities is exponential with canonical parameters

$$\theta^* = -\frac{1}{2\sigma^2} \text{ and } \lambda = \frac{nm}{\sigma^2},$$

for example. The corresponding canonical statistics are

$$T = \sum_{i=1}^{n} Y_i^2 \text{ and } S = \bar{Y}.$$

Now, consider the null hypothesis H_0: $\sigma^2 \leq \sigma_0^2$. This hypothesis bears on one canonical parameter since it can be written as

$$H_0 : \theta^* \leq \theta_0^*,$$

where $\theta_0^* = -1/2\sigma_0^2$. Thus there exists a UMPU test of this hypothesis at level α.

To determine such a test, we note that the statistic

$$\frac{U}{\sigma_0^2} = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\sigma_0^2}$$

follows a $\chi^2(n-1)$ distribution when $\sigma^2 = \sigma_0^2$. Hence U is ancillary when $\sigma^2 = \sigma_0^2$, i.e., when $\theta^* = \theta_0^*$.

Moreover, we have

$$\frac{U}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n} Y_i^2 - \frac{n}{\sigma_0^2}\bar{Y}^2 = \frac{1}{\sigma_0^2}T - \frac{n}{\sigma_0^2}S^2.$$

Thus U is a (linear) function increasing in T. Applying Property 16.5, it follows that a UMPU test at level α is given by the critical region

$$W = \left\{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \geq \sigma_0^2 \chi^2_{1-\alpha}(n-1)\right\},$$

where $\chi^2_{1-\alpha}(n-1)$ is the $(1-\alpha)$-th quantile of the chi-square distribution with $n-1$ degrees of freedom.

d) Student Tests

One of the most important applications of the theory of optimal unbiased tests deals with the linear model. It is useful to study this case separately.

We consider the linear model

$$Y = X_1 b_1 + \ldots + X_K b_K + u,$$

where Y is a vector of n observations on the endogenous variable, $X_1, \ldots, X_K$ are the vectors of observations on K nonstochastic exogenous variables, and u is a vector of random errors whose joint distribution is $N(0, \sigma^2\mathbf{I})$.

The density of such a model is given by

$$\begin{aligned}\ell(Y; b, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - \mathbf{X}b)'(Y - \mathbf{X}b)\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{b'\mathbf{X}'\mathbf{X}b}{2\sigma^2}\right) \exp\left(-\frac{\|Y\|^2}{2\sigma^2} + \frac{b'}{\sigma^2}\mathbf{X}'Y\right),\end{aligned}$$

where $\mathbf{X} = (X_1, \ldots, X_K)$ and $b = (b_1, \ldots, b_K)'$. This family is exponential with canonical parameters

$$\theta_0 = \frac{1}{\sigma^2}, \quad \theta_k = \frac{b_k}{\sigma^2}, \quad k = 1, \ldots, K.$$

In applied work one is frequently led to consider hypotheses about a linear combination of the parameters of the deterministic component. These hypotheses can be of the form

$$\tilde{H}_0 : \sum_{k=1}^{K} a_k b_k \leq a_0$$

or of the form

$$H_0 : \sum_{k=1}^{K} a_k b_k = a_0,$$

where a_k, $k = 0, \ldots, K$, are known real numbers.

Such hypotheses only involve the canonical parameters since they can be written respectively as

$$\tilde{H}_0 : \sum_{k=1}^{K} a_k \theta_k - a_0 \theta_0 \leq 0$$

and

$$H_0 : \sum_{k=1}^{K} a_k \theta_k - a_0 \theta_0 = 0.$$

Thus, from the preceding properties, there exist UMPU tests for such hypotheses.

Definition 16.5: *A one-sided Student test is a UMPU test of the null hypothesis $\tilde{H}_0$. A two-sided Student test is a UMPU test of the null hypothesis H_0.*

These tests are intuitively easy to construct. To simplify, we consider the case of testing the null hypothesis H_0: $b_1 = 0$. The resulting test is called a *test of significance of the coefficient* b_1. Because the coefficient b_1 is unknown, it is natural to estimate it by the maximum likelihood method, which corresponds here to the ordinary least squares (OLS) method.

Let $\mathbf{X}^* = (X_2, \ldots, X_k)$. Then the OLS estimator of b_1 is

$$\hat{b}_1 = \left(X_1' (\mathbf{I} - \mathbf{X}^* (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime}) X_1 \right)^{-1} X_1' (\mathbf{I} - \mathbf{X}^* (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime}) Y.$$

The OLS estimator is normally distributed with mean equal to b_1 and variance

$$\begin{aligned} V\hat{b}_1 &= \sigma^2 \left(X_1' (\mathbf{I} - \mathbf{X}^* (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime}) X_1 \right)^{-1} \\ &= \sigma^2 \omega_{11}. \end{aligned}$$

Because the null hypothesis corresponds to the nullity of b_1, it is appropriate to reject the null hypothesis when $|\hat{b}_1|$ is too large. Thus an intuitive critical region is of the form $W = \{|\hat{b}_1| > c\}$. It remains to determine the critical level c (random or nonstochastic) so as to obtain the desired risk of Type I error.

To do so, it is necessary to determine the probability distribution of $\hat{b}_1$ under the null hypothesis $b_1 = 0$. Under the null hypothesis, $\hat{b}_1/(\sigma\sqrt{\omega_{11}})$ follows a standard normal distribution. But this ratio cannot be used as a test statistic since it depends on σ^2, which is unknown. It is known, however, that the variance σ^2 can be estimated unbiasedly by

$$s^2 = \frac{1}{n-K}\|Y - X_1\hat{b}_1 - \cdots - X_K\hat{b}_K\|^2 = \frac{1}{n-K}\sum_{i=1}^{n}\hat{u}_i^2.$$

Replacing σ by $\sqrt{s^2} = s$, we obtain the ratio

$$t = \frac{\hat{b}_1}{s\sqrt{\omega_{11}}}, \tag{16.3}$$

which is a function of the observations only.

Consider now the distribution of the latter ratio under the null hypothesis $b_1 = 0$. We have

$$t = \frac{\hat{b}_1}{\sigma\sqrt{\omega_{11}}}\Big/\sqrt{\frac{s^2}{\sigma^2}}.$$

From Cochran Theorem B.6, it follows that the random variables $\hat{b}_1/(\sigma\sqrt{\omega_{11}})$ and $(n-K)s^2/\sigma^2$ are independent and distributed as a standard normal $N(0,1)$ and a chi-square distribution $\chi^2(n-K)$ respectively. Therefore, under H_0: $b_1 = 0$, the statistic t follows a Student's t-distribution with $n-K$ degrees of freedom.

Definition 16.6:

(i) The Student statistic (or t-statistic) associated with the coefficient b_1 is

$$t = \frac{\hat{b}_1}{s\sqrt{\omega_{11}}} = \frac{\hat{b}_1}{\hat{\sigma}_{\hat{b}_1}}.$$

(ii) If $b_1 = 0$ then this statistic follows a $t(n-K)$ distribution, i.e., a Student's t-distribution with $(n-K)$ degrees of freedom.

It is now easy to determine the critical level c so as to obtain the desired risk of Type I error. Specifically let $t_{1-\alpha/2}(n-K)$ denote the

$(1-\alpha/2)$-th quantile of the Student's t-distribution with $n-K$ degrees of freedom. Define the critical region as

$$W = \{|t| > t_{1-\alpha/2}(n-K)\} = \left\{|\hat{b}_1| > \hat{\sigma}_{\hat{b}_1} t_{1-\alpha/2}(n-K)\right\}. \quad (16.4)$$

The corresponding test satisfies $P_{b_1=0}(W) = \alpha$.

Note, however, that we have not yet established the optimality of the above test. This is the purpose of the next property.

Property 16.10: *The test with critical region*

$$W = \left\{|\hat{b}_1| > \hat{\sigma}_{\hat{b}_1} t_{1-\alpha/2}(n-K)\right\},$$

is UMPU at level α for testing $H_0 : b_1 = 0$ against $H_1 : b_1 \neq 0$.

PROOF: First, without loss of generality, it is possible to assume that X_1 is orthogonal to $X_2, \ldots, X_K$. This is because we have

$$\begin{aligned} X_1 b_1 + \ldots + X_K b_K &= \mathbf{M}^* X_1 b_1 + X_2 b_2^* + \ldots + X_K b_K^* \\ &= X_1^* b_1 + X_2 b_2^* + \ldots + X_K b_K^*, \end{aligned}$$

where $X_1^* = \mathbf{M}^* X_1$ denotes the orthogonal projection on to the subspace orthogonal to the space generated by $X_2, \ldots, X_K$ and $b_2^*, \ldots, b_K^*$ are defined by

$$X_2 b_2^* + \ldots + X_K b_K^* = X_2 b_2 + \ldots + X_K b_K + (\mathbf{I} - \mathbf{M}^*) X_1 b_1.$$

Then the variable $X_1^* = \mathbf{M}^* X_1$ is orthogonal to $X_2, \ldots, X_K$ by construction.

Second, given such an orthogonality, various simplifications arise in the expressions for the OLS estimator and its variance. Namely

$$\hat{b}_1 = (X_1' X_1)^{-1} X_1' Y \quad \text{and} \quad \omega_{11} = (X_1' X_1)^{-1}.$$

Moreover, the likelihood function becomes

$$\ell(Y; b, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{b'\mathbf{X}'\mathbf{X}b}{\sigma^2}\right) \exp\left(-\frac{\|Y\|^2}{2\sigma^2} + \frac{b_1}{\sigma^2 \omega_{11}} \hat{b}_1 + \frac{b^{*\prime}}{\sigma^2} \mathbf{X}^{*\prime} Y\right),$$

where $b^* = (b_2, \ldots, b_K)'$ and $\mathbf{X}^* = (X_2, \ldots, X_K)$.

Third, we saw previously that the Student statistic $t = \hat{b}_1/(s\sqrt{\omega_{11}})$ is ancillary under H_0: $b_1 = 0$, i.e., under H_0: $b_1/\sigma^2 = 0$.

In order to apply Property 16.8, it remains to study how the statistic t depends on

$$T(Y) = \hat{b}_1, \quad S(Y) = \begin{pmatrix} \|Y\|^2 \\ \mathbf{X}^{*\prime} Y \end{pmatrix} = \begin{pmatrix} S_1(Y) \\ S_2(Y) \end{pmatrix}.$$

We have

$$t = \frac{\hat{b}_1}{s\sqrt{\omega_{11}}} = \frac{\sqrt{n-K}\hat{b}_1}{\sqrt{\omega_{11}}\left(\|Y\|^2 - \|\mathbf{X}^*\hat{b}^*\|^2 - \hat{b}_1^2/\omega_{11}\right)^{1/2}},$$

where $\hat{b}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}Y$. This relation can also be written as

$$t = \frac{\sqrt{n-K}U}{(1-U^2)^{1/2}},$$

or equivalently

$$U = \frac{t}{(n-K+t^2)^{1/2}},$$

where

$$\begin{aligned} U &= \frac{\hat{b}_1}{\sqrt{\omega_{11}}} \frac{1}{(\|Y\|^2 - \|\mathbf{X}^*\hat{b}^*\|^2)^{1/2}} \\ &= \frac{T(Y)}{\sqrt{\omega_{11}}} \frac{1}{(S_1 - \|\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}S_2\|^2)^{1/2}}. \end{aligned}$$

The statistic U is in a bijective relationship with the Student statistic t. Thus the statistic U is also ancillary under $b_1 = 0$. Since U is a linear strictly increasing function of the canonical statistic $T(Y)$, we can apply Property 16.8 to the statistic U. It follows that there exists a UMPU test. Its critical region is of the form $W = \{|U| > \tilde{c}\}$. Since $|U|$ and $|t|$ are related by an increasing function, this critical region is also of the form $W = \{|t| > c\}$. Because of the constraint on the risk of Type I error, this critical region coincides with the critical region $W = \{|t| > t_{1-\alpha/2}(n-K)\}$. □

A similar argument can be used to establish the optimality of the Student test for testing the one-sided hypothesis $\tilde{H}_0$:$b_1 \leq 0$.

Property 16.11: *There exists a UMPU test for testing $\tilde{H}_0 : b_1 \leq 0$. Namely, the Student test with critical region $\tilde{W} = \{t > t_{1-\alpha}(n-K)\}$ is UMPU.*

Remark 16.4: When the number n of observations is large, the Student's t-distribution $t(n-K)$ is close to the standard normal distribution. Therefore the critical regions of Property 16.11 and Remark 16.4 can be approximated as follows

$$W \approx \{|t| > z_{1-\alpha/2}\} \text{ and } \tilde{W} \approx \{t > z_{1-\alpha}\},$$

For instance, when $\alpha = 5\%$, this gives approximately

$$W \approx \{|t| > 2\} \text{ and } \tilde{W} \approx \{t > 1.64\}.$$

Remark 16.5: When the estimation results of a linear model are presented, one frequently places the t-statistics under the corresponding estimated coefficients. Thus one knows at once whether some explanatory variables are significant or not.

Student tests can be defined similarly when the null hypothesis concerns a linear combination of the coefficients of the deterministic component. Specifically, for testing the null hypothesis H_0: $\sum_{k=1}^{K} a_k b_k = a_0$, we consider the Student statistic associated with the coefficient $\theta^* = \sum_{k=1}^{K} a_k b_k - a_0$. This statistic is given by

$$t = \frac{a'\hat{b} - a_0}{s(a'(\mathbf{X}'\mathbf{X})^{-1}a)^{1/2}}, \tag{16.5}$$

where $a = (a_1, \ldots, a_K)'$ and $\hat{b}$ is the OLS estimator of b.

The critical region of the Student test for testing H_0: $\theta^* = 0$ is

$$W = \left(|a'\hat{b} - a_0| > t_{1-\alpha/2}(n-K)s(a'(\mathbf{X}'\mathbf{X})^{-1}a)^{1/2}\right). \tag{16.6}$$

Example 16.5: Chow Test. The relation linking the endogenous variable to the explanatory variables may not be stable over time. It is then important to have some tests that can detect eventual *breaks*. For instance, consider the basic model

$$Y_t = at + b + u_t, \quad t = 1, \ldots, T, \quad u \sim N(0, \sigma^2\mathbf{I}),$$

where the endogenous variable Y is specified as a linear function of time. One may want to test whether there is a change in the growth rate at time t_0. If one assumes that the variable Y varies continuously over time, then one is led to consider the nesting model

$$\begin{cases} Y_t = \alpha_1(t - t_0) + \beta + u_t, & \text{if } t \leq t_0, \\ Y_t = \alpha_2(t - t_0) + \beta + u_t, & \text{if } t \geq t_0. \end{cases}$$

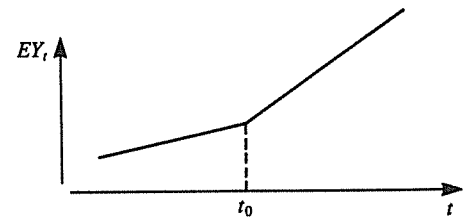


Figure 16.1: Linear Model with Different Growth Rates

The nesting model corresponds to a linear trend over each period with a change of slopes at $t = t_0$. See Figure 16.1.

The nesting model is linear because it can be written as

$$Y_t = \alpha_1(t - t_0)\mathbb{1}_{t-t_0\leq 0} + \alpha_2(t - t_0)\mathbb{1}_{t-t_0\geq 0} + \beta + u_t,$$

for $t = 1, \ldots, T$.

The hypothesis of no break can be written as H_0: $\alpha_1 = \alpha_2$. The corresponding Student test, which is called here the Chow test, has a critical region given by

$$W = \left\{|\hat{\alpha}_1 - \hat{\alpha}_2| \geq (\hat{V}\hat{\alpha}_1 + \hat{V}\hat{\alpha}_2 - 2\widehat{\text{cov}}(\hat{\alpha}_1, \hat{\alpha}_2))^{1/2} t_{1-\alpha/2}(T-3)\right\}.$$

This example shows that knowledge of the Student statistics associated with α_1 and α_2, and, in particular, of $\hat{V}\hat{\alpha}_1$ and $\hat{V}\hat{\alpha}_2$ may not be sufficient for constructing tests of some interesting hypotheses. Specifically, it may be necessary to know the estimated covariances of the coefficient estimates, namely $\widehat{\text{cov}}(\hat{\alpha}_1, \hat{\alpha}_2)$ in the present case.

e) Tests of Independence in 2×2 Contingency Tables

Another important application of the theory of UMPU tests is the analysis of contingency tables. Suppose that the frequencies n_{ij}, $i = 0, 1$ and $j = 0, 1$, in a 2×2 contingency table follow a multinomial distribution $M(n; p_{ij}, i = 0, 1, j = 0, 1)$.

$Y \rightarrow$ $X \downarrow$	0	1	Total
0	n_{00}	n_{01}	$n_{0\cdot}$
1	n_{10}	n_{11}	$n_{1\cdot}$
Total	$n_{\cdot 0}$	$n_{\cdot 1}$	n

It may be interesting to know whether the two dichotomous variables X and Y in this 2×2 table are independent. In terms of the parameters of the model, i.e., of the cell probabilities, such a hypothesis can be written as

$$H_0 : p_{00}p_{11} = p_{01}p_{10},$$

or equivalently as

$$H_0 : \log p_{00} + \log p_{11} - \log p_{01} - \log p_{10} = 0.$$

Thus the problem reduces to testing the nullity of a linear combination of the logarithms of the cell probabilities. Because the coefficients associated with this linear combination sum to zero, it follows from Example 16.3 that there exists a UMPU test at level α for testing the independence of Y and X.

To find the form of such an optimal test, we shall rewrite the density of the model so as to single out the parameter of interest

$$\theta^* = \log \frac{p_{11}p_{00}}{p_{10}p_{01}}.$$

We have

$$\begin{aligned} \ell(n;p) &= \frac{n!}{n_{00}!n_{01}!n_{10}!n_{11}!} p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}} \\ &= \frac{n!}{n_{00}!n_{01}!n_{10}!n_{11}!} \exp(n \log p_{11}) \exp(\theta^* n_{00} + \lambda_1 n_{0\cdot} + \lambda_2 n_{\cdot 0}), \end{aligned}$$

where

$$\theta^* = \log \frac{p_{11}p_{00}}{p_{10}p_{01}}, \quad \lambda_1 = \log \frac{p_{01}}{p_{11}}, \quad \lambda_2 = \log \frac{p_{10}}{p_{11}},$$

and

$$T = n_{00}, \quad S_1 = n_{0\cdot} = n_{00} + n_{01}, \quad S_2 = n_{\cdot 0} = n_{00} + n_{10}.$$

From Property 16.7 we know that a UMPU test can be constructed from the conditional distribution of T given $S = (S_1, S_2)'$. This conditional distribution is readily obtained by applying the next lemma whose proof is left as an exercise.

Lemma 16.1: *Conditionally upon $n_{\cdot 0} = n_{00} + n_{10}$, the random variables n_{00} and n_{01} are independent and follow respectively the Binomial distributions*

$$B\left(n_{\cdot 0}, \frac{p_{00}}{p_{10} + p_{00}}\right) \text{ and } B\left(n - n_{\cdot 0}, \frac{p_{01}}{p_{01} + p_{11}}\right).$$

Now, under the null hypothesis of independence, the ratios $p_{00}/(p_{10}+p_{00})$ and $p_{01}/(p_{01}+p_{11})$, which can be viewed as conditional probabilities, are equal. Then it is easy to see that the conditional distribution of $T = n_{00}$ given $S = (n_{0\cdot}, n_{\cdot 0})$ under H_0 is a hypergeometric distribution. That is

$$P_0\left(T = t \mid S = (n_{0\cdot}, n_{\cdot 0})\right) = \frac{\binom{n_{\cdot 0}}{t}\binom{n - n_{\cdot 0}}{n_{0\cdot} - t}}{\binom{n}{n_{0\cdot}}}.$$

Using a table of hypergeometric distributions, we can determine in principle the values $c_1(s)$, $c_2(s)$, $\gamma_1(s)$, and $\gamma_2(s)$ of Property 16.7 for each value of $(n_{0\cdot}, n_{\cdot 0}) = s$. Such a determination, however, is relatively complex. This is one reason for introducing asymptotic tests, which are discussed in the next chapter.

16.1.4 Locally Uniformly Most Powerful Unbiased Tests

To obtain UMPU tests in Sections 16.1.2 and 16.1.3, we have assumed that the null hypothesis relates to one parameter only and that the model is exponential. We now consider the case where the model is no longer exponential. We still assume that the model has one parameter only. The null hypothesis to be tested is H_0: $\theta = \theta_0$ against the alternative hypothesis H_1: $\theta \neq \theta_0$. Thus we shall consider a two-sided test.

The method discussed in Section 15.2.3, which is based on finding a locally most powerful test, no longer applies. For this method reduces here to finding a test ϕ that maximizes simultaneously

$$\left(\frac{\partial E_\theta \phi}{\partial \theta}\right)_{\theta=\theta_0} \text{ and } -\left(\frac{\partial E_\theta \phi}{\partial \theta}\right)_{\theta=\theta_0}$$

in the class of tests ϕ that satisfy $E_{\theta_0}\phi = \alpha$.

A possible solution is to restrict our search to tests that are locally unbiased at level α.

Definition 16.7: *A test ϕ is locally unbiased at level α if $E_\theta\phi$ is twice differentiable in θ_0 and if*

$$\begin{aligned} E_{\theta_0}\phi &= \alpha, \\ \left(\frac{\partial E_\theta \phi}{\partial \theta}\right)_{\theta=\theta_0} &= 0, \\ \left(\frac{\partial^2 E_\theta \phi}{\partial \theta^2}\right)_{\theta=\theta_0} &> 0. \end{aligned}$$

A test ϕ is locally uniformly most powerful and unbiased (LUMPU) at level α if it is locally unbiased at level α and if

$$\left(\frac{\partial^2 E_\theta \phi}{\partial \theta^2}\right)_{\theta=\theta_0} \geq \left(\frac{\partial^2 E_\theta \phi^*}{\partial \theta^2}\right)_{\theta=\theta_0}$$

for any other test ϕ^* that is locally unbiased at level α.

Thus, to find a LUMPU test at level α, we now consider the optimization problem

$$\max_\phi \left(\frac{\partial^2 E_\theta \phi}{\partial \theta^2}\right)_{\theta=\theta_0},$$

subject to

$$\begin{cases} E_{\theta_0}\phi = \alpha, \\ \left(\frac{\partial E_\theta \phi}{\partial \theta}\right)_{\theta=\theta_0} = 0. \end{cases}$$

In the parametric case where P_θ has a density $\ell(y;\theta)$ with respect to some measure μ, this optimization problem can be written as

$$\max_\phi \int \phi \frac{\partial^2 \ell(y;\theta_0)}{\partial \theta^2} d\mu,$$

subject to

$$\begin{cases} \int \phi \ell(y;\theta_0) d\mu = \alpha, \\ \int \phi (\partial \ell(y;\theta_0)/\partial\theta) d\mu = 0. \end{cases}$$

A simple extension of the Neyman-Pearson Theorem gives the solution to this problem.

Property 16.12: *The test defined by*

$$\phi = 1, \text{ if } \frac{\partial^2 \ell(y;\theta_0)}{\partial\theta} > k_1 \frac{\partial \ell(y;\theta_0)}{\partial\theta} + k_2 \ell(y;\theta_0),$$

$$\phi = 0, \text{ if } \frac{\partial^2 \ell(y;\theta_0)}{\partial\theta^2} < k_1 \frac{\partial \ell(y;\theta_0)}{\partial\theta} + k_2 \ell(y;\theta_0),$$

$$\phi = \gamma, \text{ if } \frac{\partial^2 \ell(y;\theta_0)}{\partial\theta^2} = k_1 \frac{\partial \ell(y;\theta_0)}{\partial\theta} + k_2 \ell(y;\theta_0),$$

with

$$E_{\theta_0}\phi = \alpha \text{ and } \left(\frac{\partial E_\theta \phi}{\partial\theta}\right)_{\theta=\theta_0} = 0,$$

is LUMPU at level α for testing the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

PROOF: Suppose that ϕ satisfies the stated conditions. Then, for any test ϕ^*, we have

$$\int (\phi - \phi^*) \left(\frac{\partial^2 \ell_0}{\partial\theta^2} - k_1 \frac{\partial \ell_0}{\partial\theta} - k_2 \ell_0 \right) d\mu \geq 0,$$

where $\ell_0 \equiv \ell(y;\theta_0)$.

Now, if ϕ^* is also locally unbiased at level α, then the preceding inequality becomes

$$\int (\phi - \phi^*) \frac{\partial^2 \ell_0}{\partial\theta^2} d\mu \geq 0.$$

where the constraints on ϕ have also been used. The desired result follows. □

Remark 16.6: The condition

$$\frac{\partial^2 \ell(y;\theta_0)}{\partial\theta^2} > k_1 \frac{\partial \ell(y;\theta_0)}{\partial\theta} + k_2 \ell(y;\theta_0)$$

can be written as

$$\frac{1}{\ell(y;\theta_0)} \frac{\partial^2 \ell(y;\theta_0)}{\partial\theta^2} > \frac{k_1}{\ell(y;\theta_0)} \frac{\partial \ell(y;\theta_0)}{\partial\theta} + k_2.$$

Introducing the score $S_{\theta_0}(y) = \partial \log \ell(y;\theta_0)/\partial\theta$, this condition can also be written as

$$\frac{\partial S_{\theta_0}(y)}{\partial\theta} + S^2_{\theta_0}(y) > k_1 S_{\theta_0}(y) + k_2.$$

Remark 16.7: Suppose that we consider either a sampling model, where $\ell(y;\theta) = \prod_{i=1}^n f(y_i;\theta)$ or a conditional model, where $\ell(y;\theta) = \prod_{i=1}^n f(y_i \mid x_i;\theta)$. Under appropriate regularity conditions (see Property 15.6), we know that the variable $(1/\sqrt{n})S_{\theta_0}(Y)$ is asymptotically distributed as a centered normal variable under H_0. Moreover, $(1/n)$ $(\partial S_{\theta_0}(Y)/\partial\theta)$ converges almost surely to a constant c. Therefore the critical region obtained above for a sample of size n, namely

$$\left\{\frac{\partial S_{\theta_0}(Y)}{\partial\theta} + S^2_{\theta_0}(Y) > k_1(n) S_{\theta_0}(Y) + k_2(n)\right\},$$

which can also be asymptotically written as

$$\left\{\left(\frac{S_{\theta_0}(Y)}{\sqrt{n}}\right)^2 > \frac{k_1(n)}{\sqrt{n}}\frac{S_{\theta_0}(Y)}{\sqrt{n}} + \frac{k_2(n)}{n} - c\right\},$$

is asymptotically equivalent to a critical region of the form

$$\left\{\frac{S_{\theta_0}(Y)}{\sqrt{n}} > c_1(n)\right\} \cup \left\{\frac{S_{\theta_0}(Y)}{\sqrt{n}} < c_2(n)\right\}$$

or using the conditions of Property 16.12

$$\left\{\left|\frac{S_{\theta_0}(Y)}{\sqrt{n}}\right| > c^*\right\}.$$

Then a reasoning similar to the Proof of Property 15.6 shows that, under suitable regularity conditions, this LUMPU test is asymptotically a nonrandomized test with critical region

$$\left\{|S_{\theta_0}(Y)| > \sqrt{n\hat{\mathcal{I}}(\theta_0)}z_{1-\alpha/2}\right\},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution $N(0,1)$ and $\hat{\mathcal{I}}(\theta_0)$ is a consistent estimator of the Fisher information associated with one observation (see Property 15.6). This test is called a *two-sided score test.*

For instance, in the case of a Cauchy sampling (see Example 15.4), this two-sided test has a critical region given by

$$\left\{ \left| \sum_{i=1}^{n} \frac{y_i - \theta_0}{1 + (y_i - \theta_0)^2} \right| \geq \frac{\sqrt{n}}{2\sqrt{2}} z_{1-\alpha/2} \right\}.$$

Remark 16.8: Suppose that θ is a vector of $\mathbb{R}^p$. Suppose that the null hypothesis of interest is H_0: $\theta = \theta_0$.

First, we can test this null hypothesis against the alternative hypothesis $H_1(\nu)$: $\{\theta = \theta_0 + \lambda\nu, \lambda \in \mathbb{R}\}$ for a given direction ν in $\mathbb{R}^p$. The critical region of the corresponding score test is

$$\left\{ \frac{(\nu' S_{\theta_0}(Y))^2}{n\nu' \hat{\mathcal{I}}(\theta_0)\nu} \geq \chi^2_{1-\alpha}(1) \right\},$$

where $S_{\theta_0}(Y)$ is the score vector $\partial \log \ell(Y; \theta_0)/\partial\theta$ and $\mathcal{I}(\theta_0)$ is the Fisher information matrix at θ_0.

Second, we can maximize the statistic $(\nu' S_{\theta_0}(Y))^2 / (n\nu' \hat{\mathcal{I}}(\theta_0)\nu)$ with respect to ν. This problem can be written as

$$\max_{\delta} \frac{(\delta' \hat{\mathcal{I}}^{-1/2}(\theta_0) S_{\theta_0}(Y))^2}{\delta'\delta}$$

where $\delta = \hat{\mathcal{I}}^{1/2}(\theta_0)\nu$.

Without loss of generality, we can impose the constraint $\delta'\delta = 1$. Then the problem reduces to

$$\begin{cases} \max_{\delta} (\delta' \hat{\mathcal{I}}^{-1/2}(\theta_0) S_{\theta_0}(Y))^2 \\ \\ \text{subject to } \delta'\delta = 1. \end{cases}$$

The maximum is attained when

$$\delta = \frac{\hat{\mathcal{I}}^{-1/2}(\theta_0) S_{\theta_0}(Y)}{\|\hat{\mathcal{I}}^{-1/2}(\theta_0) S_{\theta_0}(Y)\|}.$$

Therefore the maximum of the above statistic is

$$\frac{1}{n} S'_{\theta_0}(Y) \hat{\mathcal{I}}^{-1}(\theta_0) S_{\theta_0}(Y).$$

This is the *general score statistic*, which will be studied in the next chapter.

16.2 Invariant Tests

The concept of unbiased tests allows us to establish some optimality results when the null hypothesis bears on one parameter. The remaining parameters possibly appearing in the model are then considered as nuisance parameters. When the null hypothesis involves many parameters simultaneously, the theory of unbiased tests is no longer appropriate. It is, however, possible to invoke another principle that can be used to reduce the set of considered tests. This principle is called the *invariance principle.*

16.2.1 Models and Problems With Invariant Tests

Definition 16.8: *The statistical model* $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$ *is invariant under a group* $\mathcal{S}$ *of bijections from* $\mathcal{Y}$ *onto itself if, for every bijection* s *in* $\mathcal{S}$ *and every* $\theta \in \Theta$, *the image of* P_θ *by* s *is a probability distribution* $P_{\theta'}$ *belonging to the family* $\{P_\theta, \theta \in \Theta\}$.

When the model is identified, the parameter value θ' is uniquely defined by θ and the mapping s. Thus, to every $s \in \mathcal{S}$, we can associate a mapping $\bar{s}$ from Θ onto itself such that $\theta' = \bar{s}(\theta)$. We let $\bar{\mathcal{S}}$ denote the set of these mappings $\bar{s}$. Because $\mathcal{S}$ is a group of bijections, it is easy to verify that $\bar{\mathcal{S}}$ is also a group of bijections.

Definition 16.9: *Consider a model* $(\mathcal{Y}, \{P_\theta, \theta \in \Theta\})$ *and a testing problem defined by the null hypothesis* $H_0 : \theta \in \Theta_0$. *A testing problem is invariant under a group* $\mathcal{S}$ *if:*

(i) The model is invariant under $\mathcal{S}$ *and*

(ii) The null hypothesis is invariant under $\bar{\mathcal{S}}$, *i.e.*

$$\forall \bar{s} \in \bar{\mathcal{S}}, \bar{s}(\Theta_0) = \Theta_0.$$

Note that condition (ii) is equivalent to

$$\bar{s}(\Theta_1) = \Theta_1, \ \forall \bar{s} \in \bar{\mathcal{S}}$$

since $\bar{s}(\Theta_1) = \bar{s}({}^c\Theta_0) = {}^c\bar{s}(\Theta_0)$.

16.2.2 Invariant and Maximal Invariant Tests

When the testing problem is invariant under $\mathcal{S}$, the invariance principle restricts the class of tests to the class of invariant tests, i.e., to tests that satisfy the following definition.

Definition 16.10: *A test ϕ is invariant under $\mathcal{S}$ if*

$$\phi(y) = \phi(s(y)), \quad \forall y \in \mathcal{Y}, \forall s \in \mathcal{S}.$$

A rationale can be easily given to the invariance principle. For, if we consider the model that is induced by the bijection $s \in \mathcal{S}$, then the testing problem remains formally identical to the original testing problem since the specified family of probability distributions and the null hypothesis do not change. Thus it is natural to require that the test ϕ gives the same answer whether it is applied to the observations y in the original model or to the transformed observations $s(y)$.

In this section, we shall invoke this principle, i.e., we shall restrict ourselves to invariant tests, and then we shall apply the Neyman principle to this restricted set of tests. In particular, we shall investigate whether there exists a uniformly most powerful test in the class of invariant tests. Such a test, when it exists, is called a *uniformly most powerful invariant* (UMPI) test.

The search for UMPI tests is based on the concepts of invariant functions and maximal invariant functions.

Definition 16.11: *A function V defined on $\mathcal{Y}$ is invariant under $\mathcal{S}$ if*

$$V(y) = V(s(y)), \quad \forall y \in \mathcal{Y} \text{ and } \forall s \in \mathcal{S}$$

A function W defined on $\mathcal{Y}$ is maximal invariant under $\mathcal{S}$ if

$$W(y) = W(y') \Leftrightarrow \exists s \in \mathcal{S},\ y' = s(y).$$

Similarly, it is possible to define an invariant or a maximal invariant function defined on Θ under $\bar{\mathcal{S}}$.

It is clear that a maximal invariant function is also invariant. Moreover, if $W(y) = W(y')$, then $V(y) = V(y')$ for every invariant function V. Thus every invariant function V depends on y through W, namely $V(y) = h(W(y))$ for some function h. In particular, if there exists a maximal invariant statistic $T(y)$, then any invariant tests ϕ can be written as $\phi(y) = \psi(T(y)$ for some function ψ.

Hence the invariance principle leads naturally to the study of the model induced by T. The next result shows that the parameter space of this induced model can be reduced in general.

Property 16.13: *Let V be an invariant statistic under $\mathcal{S}$ and let u be a function defined on Θ that is maximal invariant under $\bar{\mathcal{S}}$. Then the probability distribution of $V(Y)$ depends on θ only through $u(\theta)$.*

PROOF: Let ν be an arbitrary subset of possible values of V and s be an arbitrary element of $\mathcal{S}$. We have

$$\begin{aligned} P_\theta(V(Y) \in \nu) &= P_\theta(V(s(Y)) \in \nu) \\ &= P_\theta(s(Y) \in V^{-1}(\nu)) \\ &= P_{\bar{s}(\theta)}(Y \in V^{-1}(\nu)) \\ &= P_{\bar{s}(\theta)}(V(Y) \in \nu). \end{aligned}$$

Thus the mapping $\theta \to P_\theta(V(Y) \in \nu)$ is invariant under $\bar{\mathcal{S}}$. Because u is maximal invariant under $\bar{\mathcal{S}}$, it follows that the mapping $\theta \to P_\theta(V(Y) \in \nu)$ depends on θ only through $u(\theta)$. Because ν is arbitrary, the desired result follows. □

This reduction of the parameter space is used frequently to transform the null hypothesis H_0 into a hypothesis about one scalar parameter (possibly in the presence of nuisance parameters) so as to apply the results derived in the previous sections.

16.2.3 Application: Fisher Tests

We consider the linear model

$$Y = \mathbf{X}b + u,$$

where Y is an observed random vector with values in $\mathbb{R}^n$, $\mathbf{X}$ is a non-stochastic $n \times K$ matrix with rank equal to K, b is an unknown fixed vector of parameters with values in $\mathbb{R}^K$, and u is an unobserved vector of random disturbances with values in $\mathbb{R}^n$ whose joint distribution is $N(0, \sigma^2\mathbf{I})$ with σ^2 unknown.

We want to test the *linear hypothesis*

$$H_0 : \mathbf{G}b = 0, \tag{16.7}$$

where $\mathbf{G}$ is a given matrix of dimension $G \times K$ and of rank requal to G.

Let $\mathcal{L} = \{\mathbf{X}b, b \in I\!R^K\}$ be the subspace of $I\!R^n$ of dimension K that is generated by the column vectors of $\mathbf{X}$. Let $\mathcal{L}_0 = \{\mathbf{X}b, b \in I\!R^K, \mathbf{G}b = 0\}$ be the subspace of $I\!R^n$ included in $\mathcal{L}$ of dimension $K - G$ that contains the mean of Y under H_0. We denote by $\mathcal{L}_1$ the subspace of dimension G that is orthogonal to $\mathcal{L}_0$ in $\mathcal{L}$ and by $\mathcal{L}^\perp$ the subspace of dimension $n - K$ that is orthogonal to $\mathcal{L}$ in $I\!R^n$.

First Reduction by Invariance

Let s_c denote the translation of a vector in $\mathcal{L}_0$ by the vector c, i.e., $s_c(y) = y + c$. The set of such translations is a group $\mathcal{S}_1$ under which the testing problem is invariant. Relative to this group $\mathcal{S}_1$, a maximal invariant statistic is the orthogonal projection of Y on to the subspace orthogonal to $\mathcal{L}_0$ in $I\!R^n$, i.e., on $\mathcal{L}_1 + \mathcal{L}^\perp$.

Because $\mathcal{L}_1$ and $\mathcal{L}^\perp$ are orthogonal to each other, the preceding statistic is equal to $(\mathbf{P}_{\mathcal{L}_1} + \mathbf{P}_{\mathcal{L}^\perp})Y$, where $\mathbf{P}_{\mathcal{L}_1}$ and $\mathbf{P}_{\mathcal{L}^\perp}$ denote the orthogonal projections matrices on to $\mathcal{L}_1$ and $\mathcal{L}^\perp$ respectively. Thus a test invariant relative to $\mathcal{S}_1$ depends on Y through $T_1(Y) = (\mathbf{P}_{\mathcal{L}_1} + \mathbf{P}_{\mathcal{L}^\perp})Y$. The invariance principle leads to considering the model induced by T_1.

Second Reduction by Invariance

Because $T_1(Y)$ is obtained from Y through a linear mapping, we have

$$Y_1 = T_1(Y) = (\mathbf{P}_{\mathcal{L}_1} + \mathbf{P}_{\mathcal{L}^\perp})(\mathbf{X}b + u) = \mathbf{P}_{\mathcal{L}_1}\mathbf{X}b + u_1,$$

where $u_1 \sim N(0, \sigma^2(\mathbf{P}_{\mathcal{L}_1} + \mathbf{P}_{\mathcal{L}^\perp}))$.

The probability distribution of $T_1(Y)$ is a singular normal distribution whose density with respect to the Lebesgue measure on $\mathcal{L}_1 + \mathcal{L}^\perp$ is

$$\begin{aligned}
&\ell(y_1; b, \sigma^2) \\
&= (2\pi\sigma^2)^{-\frac{n-K+G}{2}} \exp\left(-\frac{1}{2\sigma^2}(y_1 - \mathbf{P}_{\mathcal{L}_1}\mathbf{X}b)(\mathbf{P}_{\mathcal{L}_1} + \mathbf{P}_{\mathcal{L}^\perp})(y_1 - \mathbf{P}_{\mathcal{L}_1}\mathbf{X}b)\right) \\
&= (2\pi\sigma^2)^{-\frac{n-K+G}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}_{\mathcal{L}_1}y_1 + \mathbf{P}_{\mathcal{L}^\perp}y_1 - \mathbf{P}_{\mathcal{L}_1}\mathbf{X}b)\|^2\right) \\
&= (2\pi\sigma^2)^{-\frac{n-K+G}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}_{\mathcal{L}_1}y_1 - \mathbf{P}_{\mathcal{L}_1}\mathbf{X}b)\|^2 - \frac{1}{2\sigma^2}\|\mathbf{P}_{\mathcal{L}^\perp}y_1\|^2\right)
\end{aligned}$$

(see equation B.64).

The probability distributions belonging to the null hypothesis H_0 are obtained from the preceding probability distributions by setting $\mathbf{P}_{\mathcal{L}_1}\mathbf{X}b = 0$.

Consider now the mappings s_η from $\mathcal{L}_1 + \mathcal{L}^\perp$ on to itself defined as follows. If y_1 is an element of $\mathcal{L}_1 + \mathcal{L}^\perp$, then $y_1 = y_1^* + y_1^{**}$ where $y_1^* \in \mathcal{L}_1$ and $y_1^{**} \in \mathcal{L}^\perp$. Let η be an orthogonal mapping from $\mathcal{L}_1$ on to $\mathcal{L}_1$. We define

$$s_\eta(y_1) = \eta(y_1^*) + y_1^{**}.$$

When η varies in the group of orthogonal mappings of $\mathcal{L}_1$, then the mappings s_η thus defined constitute a group $\mathcal{S}_2$ of bijections from $\mathcal{L}_1 + \mathcal{L}^\perp$ on to $\mathcal{L}_1 + \mathcal{L}^\perp$.

The testing problem is invariant under this group. Moreover, a maximal invariant statistic under this group is

$$T_2(Y) = (\|\mathbf{P}_{\mathcal{L}_1}Y\|^2, \mathbf{P}_{\mathcal{L}^\perp}Y).$$

Hence an invariant test under $\mathcal{S}_2$ depends on Y only through $T_2(Y)$.

Reduction by Sufficiency

We now have to determine the probability distribution of $T_2(Y)$. Because the subspaces $\mathcal{L}_1$ and $\mathcal{L}^\perp$ are mutually orthogonal, it follows from the Cochran Theorem that the statistics $\|\mathbf{P}_{\mathcal{L}_1}Y\|^2$ and $\mathbf{P}_{\mathcal{L}^\perp}Y$ are mutually independent. Their marginal distributions are such that $\|\mathbf{P}_{\mathcal{L}_1}Y\|^2/\sigma^2$ follows a χ^2 distribution with G degrees of freedom and noncentrality parameter $\lambda = \|\mathbf{P}_{\mathcal{L}_1}\mathbf{X}b\|^2/\sigma^2$ while $\mathbf{P}_{\mathcal{L}^\perp}Y$ follows a normal distribution with density

$$(2\pi\sigma^2)^{-\frac{n-K}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{P}_{\mathcal{L}^\perp}y\|^2\right)$$

with respect to the Lebesgue measure on $\mathcal{L}^\perp$.

Note that the probability distribution of $T_2(Y)$ depends on the parameters (b, σ^2) through (λ, σ^2). Thus the latter pair is maximal invariant relative to

$$\bar{\mathcal{S}}_2 \circ \bar{\mathcal{S}}_1 = \{s_2 o s_1; s_2 \in \bar{\mathcal{S}}_2, s_1 \in \bar{\mathcal{S}}_1\}$$

From the Factorization Criterion, the statistic

$$T_3(Y) = (\|\mathbf{P}_{\mathcal{L}_1}Y\|^2, \|\mathbf{P}_{\mathcal{L}^\perp}Y\|^2),$$

is sufficient for (λ, σ^2). Thus we can consider the model induced by T_3.

Third Reduction by Invariance

The probability distribution of $T_3(Y)$ is characterized by the marginal distributions of $\| \mathbf{P}_{\mathcal{L}_1} Y \|^2$ and $\| \mathbf{P}_{\mathcal{L}^\perp} Y \|^2$, which are given by

$$\frac{\| \mathbf{P}_{\mathcal{L}_1} Y \|^2}{\sigma^2} \sim \chi^2(G, \lambda)$$

and

$$\frac{\| \mathbf{P}_{\mathcal{L}^\perp} Y \|^2}{\sigma^2} \sim \chi^2(n-K).$$

In the model induced by T_3, the null hypothesis can be written as H_0: $\lambda = 0$.

On the other hand, the scalar transformations of $\mathbb{R}^2_+$ with positive scalars constitute a group of transformations $\mathcal{S}_3$ under which the testing problem is invariant. A maximal invariant statistic under this group is

$$F = \frac{\|\mathbf{P}_{\mathcal{L}_1} Y\|^2/G}{\|\mathbf{P}_{\mathcal{L}^\perp} Y\|^2/(n-K)}.$$

This statistic follows a Fisher–Snedecor distribution with noncentrality parameter λ, namely a $F(G, n-K; \lambda)$.

Definition 16.12: *The ratio*

$$F = \frac{\|\mathbf{P}_{\mathcal{L}_1} Y\|^2/G}{\|\mathbf{P}_{\mathcal{L}^\perp} Y\|^2/(n-K)} = \frac{(\|\mathbf{P}_{\mathcal{L}} Y\|^2 - \|\mathbf{P}_{\mathcal{L}_0} Y\|^2)/G}{\|Y - \mathbf{P}_{\mathcal{L}} Y\|^2/(n-K)}.$$

is called a Fisher statistic.

This statistic follows a Fisher–Snedecor distribution

$$F(G, n-K; \lambda) \quad \text{with } \lambda = \frac{1}{\sigma^2} \|\mathbf{P}_{\mathcal{L}_1} \mathbf{X} b\|^2.$$

A UMPI Test

In the model induced by the statistic F, the probability distributions only depend on the scalar parameter $\lambda \in \mathbb{R}^+$. The null hypothesis can be written as H_0: $\lambda = 0$. It can be shown that the family $\{F(G, n-K; \lambda), \lambda \in \mathbb{R}^+\}$ of the Fisher–Snedecor distributions is a family with the monotone likelihood ratio property in the identity statistic. Therefore Property 15.3 applies and we obtain a UMP test at level α in the model induced by F. This test corresponds to a critical region of the form

$W = \{F > c\}$. The constant c is determined by the condition on the probability of Type I error, i.e., by $\mathbf{P}_{\lambda=0}(W) = \alpha$. This gives

$$c = F_{1-\alpha}(G, n-K),$$

which is $(1-\alpha)$-th quantile of the Fisher–Snedecor distribution with G and $n-K$ degrees of freedom. Thus we have proved the following result.

Property 16.14: *Consider the linear model*

$$Y = \mathbf{X}b + u, \quad u \sim N(0, \sigma^2\mathbf{I}).$$

There exists a UMPI test at level α for testing the linear hypothesis H_0: $\mathbf{G}b = 0$, where the rank of $\mathbf{G}$ is G.

The critical region of this test, which is called a Fisher test, is

$$W = \left\{ y : F = \frac{\|\mathbf{P}_{\mathcal{L}_1} y\|^2/G}{\|\mathbf{P}_{\mathcal{L}^\perp} y\|^2/(n-K)} > F_{1-\alpha}(G, n-K) \right\},$$

where n is the dimension of Y, K is the rank of $\mathbf{X}$, $\mathcal{L}$ is the subspace of $\mathbb{R}^n$ generated by the columns of $\mathbf{X}$, $\mathcal{L}_0$ is the subspace $\{\mathbf{X}b,\ b \in \mathbb{R}^K, \mathbf{G}b = 0\}$, and $\mathcal{L}_1$ is the subspace of $\mathcal{L}$ orthogonal to $\mathcal{L}_0$.

Remark 16.9: Without loss of generality, consider the linear model

$$Y = X_1 b_1 + \mathbf{X}^* b^* + u, \quad u \sim N(0; \sigma^2\mathbf{I}),$$

where X_1 is a vector orthogonal to the columns of $\mathbf{X}^*$. The Fisher statistic associated with the null hypothesis H_0: $b_1 = 0$ is

$$F = \frac{\|\mathbf{P}_{\mathcal{L}_1} Y\|^2}{\|Y - \mathbf{P}_{\mathcal{L}} Y\|^2/(n-K)} = \frac{\|X_1(X_1'X_1)^{-1}X_1'Y\|^2}{\|\hat{u}\|^2/(n-K)},$$

where $\hat{u} = Y - \mathbf{P}_{\mathcal{L}} Y$ denotes the vector of estimated residuals. Introducing the OLS estimator $\hat{b}_1 = (X_1'X_1)^{-1}X_1'Y$ of b_1, we obtain

$$F = \frac{\|X_1\hat{b}_1\|^2}{\|\hat{u}\|^2/(n-K)} = \frac{|\hat{b}_1|^2}{s^2(X_1'X_1)^{-1}} = \frac{|\hat{b}_1|^2}{s^2\omega_{11}} = t^2.$$

Thus the Fisher statistic is equal to the square of the Student statistic (see Definition 16.6).

Noting that the distribution $F(1, n-K)$ is the distribution of the square of a random variable following a Student's t-distribution with

$n - K$ degrees of freedom, i.e., a $t(n-K)$, it follows that the critical region of the Fisher test can be written as

$$\begin{aligned} W &= \{F > F_{1-\alpha}(1, n-K)\} \\ &= \{t^2 > t^2_{1-\alpha}(1, n-K)\} \\ &= \{|t| > t_{1-\alpha}(1, n-K)\}. \end{aligned}$$

Thus W is identical to the critical region associated with the Student test.

In particular, it follows that the Student test, which is optimal in the class of unbiased tests, is also optimal in the class of invariant tests. On the other hand, when more than one constraints are under test, i.e., when $G > 1$, it can be shown that there does not exist a UMPU test. Hence the Fisher test cannot be UMPU.

Remark 16.10: Let $\|\hat{u}\|^2 = \|\mathbf{P}_{\mathcal{L}^\perp}(Y)\|^2$ be the sum of squared residuals in the unconstrained linear model. Let $\|\hat{u}_0\|^2 = \|\mathbf{P}_{\mathcal{L}_0^\perp}(Y)\|^2$ be the sum of squared residuals under H_0. These are sometimes called the unconstrained and constrained sum of squares respectively. Then the Fisher statistic F can be written as

$$F = \frac{(\|\hat{u}_0\|^2 - \|\hat{u}\|^2)/G}{\|\hat{u}\|^2/(n-K)}.$$

16.3 Exercises

EXERCISE 16.1: The random pair (T, S) is assumed to have the joint density $\ell(t, s; \theta^*, \lambda) = C(\theta^*, \lambda) \exp(\theta^* t + \lambda s)$, where $\theta^{*'}$ and λ' are two parameter vectors that do not appear in the implicit dominating measure.

a) Find the marginal distribution of S.

b) Find the conditional distribution of T given S. Does it depend on λ?

EXERCISE 16.2: Consider the exponential model with densities of the form

$$\ell(y; \theta) = C(\theta) h(y) \exp\left(\theta^* T(y) + \sum_{i=1}^{p-1} \lambda_i S_i(y)\right).$$

Show that there exists a UMPU test at level α for testing the null hypothesis H_0: $\theta^* \geq \theta_0^*$ against the alternative hypothesis H_1: $\theta^* < \theta_0^*$. Hint: Use Property 16.4 after having reparameterized the model from θ^* to $-\theta^*$.

EXERCISE 16.3: Consider the model

$$Y = X_1 b_1 + \mathbf{X}^* b^* + u,$$

where $u \sim N(0, \sigma^2 \mathbf{I})$ and

$$(\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime})Y = (\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime})X_1 b_1 + u,$$

where $u \sim N(0, \sigma^2 \mathbf{I})$. It is well known that the OLS estimators of b_1 and their variance covariance matrices in these two models are identical. What can be said about the Student statistics for testing H_0: $b_1 = 0$ in these two models?

EXERCISE 16.4: Consider two random samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ drawn independently from a Poisson distribution with parameter λ and a Poisson distribution with parameter μ respectively. One wants to test the null hypothesis H_0: $\mu \leq \lambda$. Show that there exists a UMPU test at level α. In particular, using an appropriate conditioning, show that one can consider a test of the parameter of a Binomial distribution.

EXERCISE 16.5: Consider two random samples $(X_1, \ldots, X_d)$ and $(Y_1, \ldots, Y_p)$ drawn independently from the normal distributions $N(m_x, \sigma_x^2)$ and $N(m_y, \sigma_y^2)$ respectively. One wants to test the null hypothesis H_0: $\sigma_y^2 \leq a\sigma_x^2$, where a is a given positive constant.

Consider the test whose critical region is

$$\left\{ \frac{\sum_{j=1}^{p}(Y_j - \bar{Y})^2/a(p-1)}{\sum_{j=1}^{d}(X_j - \bar{X})^2/(d-1)} \leq F_{1-\alpha}(p-1, d-1) \right\},$$

where $F_{1-\alpha}(p-1, d-1)$ is the $(1-\alpha)$-th quantile of the Fisher–Snedecor distribution with $p-1$ and $d-1$ degrees of freedom. Show that this test is UMPU at level α.

EXERCISE 16.6: Consider the framework of Exercise 16.5. Assume that $\sigma_x^2 = \sigma_y^2$. One wants to test the null hypothesis H_0: $m_y \leq m_x$. Show that there exists a UMPU test at level α based on the statistic

$$U = \frac{\bar{Y} - \bar{X}}{\left(\sum_{j=1}^{p}(Y_j - \bar{Y})^2 + \sum_{j=1}^{d}(X_j - \bar{X})^2\right)^{1/2}}.$$

Determine explicitly such a test.

EXERCISE 16.7: Consider the framework of Exercise 16.6. One wants to test the null hypothesis H_0: $m_y = m_x$. Show that there exists a UMPU test at level α based on $|U|$. Determine explicitly such a test.

EXERCISE 16.8: Consider a random sample $(Y_1, \ldots, Y_n)$ drawn from the normal distribution $N(m, \sigma^2)$. One wants to test the null hypothesis H_0: $\sigma \geq \sigma_0$, where σ_0 is given. Show that there exists a UMPI test at level α relative to the group of translations in $\mathbb{R}^n$ that are parallel to the vector $(1, 1, \ldots, 1)$. Show that the critical region of the test is

$$W = \left\{ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \leq \sigma_0^2 \chi^2_{1-\alpha}(n-1) \right\}.$$

Verify that this test is also UMPU at level α.

EXERCISE 16.9: Consider the random vector $(Y_1, \ldots, Y_n)$ drawn from a multivariate distribution with density

$$(1-\nu) f_0(y_1 - \mu, \ldots, y_n - \mu) + \nu f_1(y_1 - \mu, \ldots, y_n - \mu)$$

with respect to the Lebesgue measure on $\mathbb{R}^n$, where $\mu \in \mathbb{R}$, $\nu \in \{0, 1\}$, and f_0 and f_1 are two given densities.

One wants to test the null hypothesis H_0: $\nu = 0$, i.e., that the true density is f_0. Show that there exists a UMPI test at level α relative to the group of translations on $\mathbb{R}^n$ that are parallel to the vector $(1, 1, \ldots, 1)$. Show that this test is based on the statistic

$$\frac{\int_{\mathbb{R}} f_1(y_1 + z, \ldots, y_n + z) dz}{\int_{\mathbb{R}} f_0(y_1 + z, \ldots, y_n + z) dz}.$$

16.4 References

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, Wiley.

Monfort, A. (1982). *Cours de Statistique Mathématique*, Economica.

CHAPTER 17

Likelihood Based Tests

17.1 Wald, Score, and Likelihood Ratio Tests: Principles

The optimality results discussed in the preceding two chapters are not always applicable. In particular, they are not applicable when the parameter space Θ and the subspace Θ_0 defining the null hypothesis H_0 are multidimensional. In this chapter we shall present some testing methods whose range of applications is very broad but whose main properties are asymptotic.

The three fundamental tests in the theory of asymptotic tests are the *Wald, score* (or *Lagrange multiplier*), and *likelihood ratio tests.* These tests, which are sometimes called the classical tests are particularly useful when the parameter space is multidimensional. To have an intuitive understanding of these tests, however, in this introduction we begin briefly with the simple case, where the density $\ell(y;\theta)$ of the model depends on a scalar parameter θ, and where the null hypothesis is of the form H_0: $\theta = \theta_0$ with θ_0 given.

Let $L_n(\theta)$ be the log-likelihood of the model and let $\hat{\theta}_n$ be the maximum likelihood estimator of θ. Then it is intuitive to accept H_0 if $\hat{\theta}_n$ is close to θ_0. This idea is the basis of the Wald test, which considers the distance between $\hat{\theta}_n$ and θ_0.

Similarly, a situation that is favorable to H_0 arises when the score $\partial L_n(\theta_0)/\partial\theta$ is close to zero, see Figure 17.1. In the present case the score is identical to the Lagrange multiplier associated with the constraint $\theta = \theta_0$ in the optimization problem

$$\begin{cases} \max_{\theta\in\Theta} L_n(\theta) \\ \text{subject to } \theta = \theta_0. \end{cases}$$

We can then construct a test based on how far the score or the Lagrange

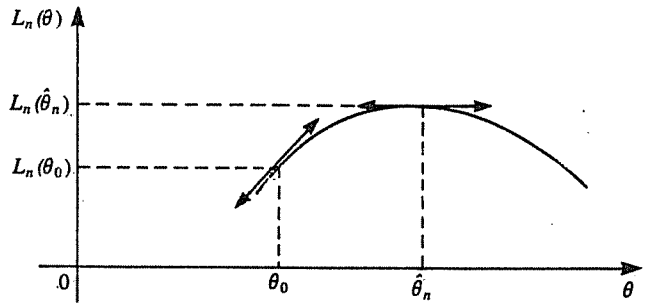


Figure 17.1: Log–likelihood Function

multiplier is from zero. This is the main idea underlying the score or Lagrange multiplier test.

Lastly, it is natural to accept H_0 when $L_n(\hat{\theta}_n)$ is close to $L_n(\theta_0)$. The likelihood ratio test exploits such an idea by relying upon the statistic

$$L_n(\hat{\theta}_n) - L_n(\theta_0) = \log \frac{\ell(y;\hat{\theta}_n)}{\ell(y;\theta_0)}.$$

17.2 Wald, Score, and Likelihood Ratio Tests When the Null Hypothesis Is of the Form $g(\theta) = 0$ (Implicit Form)

We assume that the observed variables can be partitioned into the endogenous variables Y and the exogenous variables X. The observations are (Y_i, X_i), $i = 1, \ldots, n$. To simplify, we assume that the pairs (Y_i, X_i)'s are independently and identically distributed. We are interested in the parameter vector $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^p$, and θ appears in the conditional distribution of Y_i given X_i. We let $f(y_i \mid x_i; \theta)$ denote the conditional density.

It is possible to consider more general assumptions implying the convergence of the empirical distribution of the x_i's to a limiting distribution (see Chapter 24). Within such a more general framework, the next results hold provided the joint distribution of (Y_i, X_i) in the various mathematical expectations appearing hereafter are replaced by the joint distribution whose conditional distribution is $f(y_i \mid x_i; \theta)$ and whose

marginal distribution for X_i is the limiting distribution of the empirical distribution of the x_i's.

The likelihood function used for estimating θ is the conditional likelihood function

$$\prod_{i=1}^{n} f(y_i \mid x_i; \theta).$$

We let $L_n^c(\theta)$ or $L_n(\theta)$ denote the conditional log-likelihood, i.e.

$$L_n(\theta) = \sum_{i=1}^{n} \log f(y_i \mid x_i; \theta).$$

We let $\hat{\theta}_n$ denote the (conditional) maximum likelihood estimator of θ. We also assume that the (conditional) information matrix

$$\mathcal{I}(\theta) = E_x E_\theta \left(-\frac{\partial^2 \log f(Y \mid X; \theta)}{\partial \theta \, \partial \theta'} \right)$$

is nonsingular and continuous for every $\theta \in \Theta$. Note that the expectation is taken with respect to the joint distribution of the pair (Y_i, X_i). Thus $\mathcal{I}(\theta)$ depends on the marginal distribution of X_i. Consistent estimators of $\mathcal{I}(\theta)$ are

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i \mid x_i; \theta)}{\partial \theta \, \partial \theta'} = -\frac{1}{n} \frac{\partial^2 L_n(\hat{\theta}_n)}{\partial \theta \, \partial \theta'}$$

or

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid x_i; \hat{\theta}_n)}{\partial \theta} \frac{\partial \log f(y_i \mid x_i; \hat{\theta}_n)}{\partial \theta'}.$$

We want to test the null hypothesis H_0 defined by the subset Θ_0 of values for the parameter θ satisfying

$$g(\theta) = 0 \Leftrightarrow \begin{cases} g_1(\theta) = 0, \\ \vdots \\ g_r(\theta) = 0. \end{cases} \tag{17.1}$$

The functions $g_1, \ldots, g_r$ take their values in $\mathbb{R}$, are differentiable, and are such that the matrix $\partial g'(\theta)/\partial \theta$ is of rank r, $\forall \theta \in \Theta$. The latter condition means that there are no redundant constraints. The form (17.1) of the null hypothesis is said to be *implicit*.

We consider asymptotic tests. More precisely, we consider asymptotic tests whose critical regions W_n satisfy two conditions. These conditions are defined next.

Definition 17.1: *An asymptotic test with critical region W_n has an asymptotic level equal to α if*

$$\sup_{\theta \in \Theta_0} \lim_{n \to \infty} P_\theta(W_n) = \alpha.$$

Definition 17.2: *An asymptotic test with critical region W_n is consistent if*

$$\forall \theta \in {}^c\Theta_0, P_\theta(W_n) \to 1.$$

Definition 17.1 requires that the probability of Type I error is asymptotically equal to α. On the other hand, Definition 17.2 says that the probability of Type II error is asymptotically equal to zero.

17.2.1 Wald Test

The principle of this test is to examine whether the constraint $g(\theta) = 0$ is approximately satisfied by $\hat{\theta}_n$, i.e., whether $g(\hat{\theta}_n)$ is close to 0.

Property 17.1: *Under the assumptions of Property 7.17, the test defined by the critical region*

$$\{\xi_n^W \geq \chi^2_{1-\alpha}(r)\},$$

where

$$\xi_n^W = n g'(\hat{\theta}_n) \left(\frac{\partial g(\hat{\theta}_n)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n)^{-1} \frac{\partial g'(\hat{\theta}_n)}{\partial \theta} \right)^{-1} g(\hat{\theta}_n),$$

has asymptotic level α and is consistent. This test is called the Wald test.

PROOF: Let θ_0 be the true value of θ. We know that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1}).$$

Thus from a Taylor series expansion around θ_0 we have

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta_0)\right) \xrightarrow{d} N\left(0, \frac{\partial g(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial g'(\theta_0)}{\partial \theta}\right). \tag{17.2}$$

Now, under H_0, we have $g(\theta_0) = 0$. Hence

$$\sqrt{n}g(\hat{\theta}_n) \xrightarrow{d} N\left(0, \frac{\partial g(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial g'(\theta_0)}{\partial \theta}\right).$$

Because the matrix $\partial g'(\theta_0)/\partial\theta$ is full-column rank, it follows that

$$\sqrt{n}\left(\frac{\partial g(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1/2} g(\hat{\theta}_n) \xrightarrow{d} N(0, \mathbf{I}).$$

Since $\hat{\theta}_n$ is a consistent estimator of θ_0, we have

$$\sqrt{n}\left(\frac{\partial g(\hat{\theta}_n)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n)^{-1} \frac{\partial g'(\hat{\theta}_n)}{\partial \theta}\right)^{-1/2} g(\hat{\theta}_n) \xrightarrow{d} N(0, \mathbf{I}).$$

Taking the corresponding quadratic form (see Theorem B.6), we obtain

$$\xi_n^W = n g'(\hat{\theta}_n)\left(\frac{\partial g(\hat{\theta}_n)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n)^{-1} \frac{\partial g'(\hat{\theta}_n)}{\partial \theta}\right)^{-1/2} g(\hat{\theta}_n) \xrightarrow{d} \chi^2(r)$$

under the null hypothesis $H_0 : g(\theta) = 0$. Therefore the test with the critical region defined in Property 17.1 has an asymptotic level equal to α.

This test is also consistent. Because if $\theta_0 \notin \Theta_0 = \{\theta : g(\theta) = 0\}$ then the estimator $\hat{\theta}_n$ converges to θ_0 and $g(\hat{\theta}_n)$ converges to $g(\theta_0) \neq 0$. Hence ξ_n^W converges to $+\infty$. Therefore we have

$$P_\theta\left(\xi_n^W \geq \chi^2_{1-\alpha}(r)\right) \to 1.$$

□

Recalling that the Fisher information matrix $\mathcal{I}(\theta_0)$ can be estimated in different ways, we can replace in the expression for ξ_n^W, $\mathcal{I}(\hat{\theta}_n)$ by

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i \mid x_i; \hat{\theta}_n)}{\partial \theta\, \partial \theta'} = -\frac{1}{n}\frac{\partial^2 L_n(\hat{\theta}_n)}{\partial \theta\, \partial \theta'}$$

or by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(y_i \mid x_i; \hat{\theta}_n)}{\partial \theta} \frac{\partial \log f(y_i \mid x_i; \hat{\theta}_n)}{\partial \theta'}.$$

This remark is especially important, when the conditional likelihood is used, as is the case here, since $\mathcal{I}(\theta_0)$ depends on the marginal distribution of X_i, which is unknown in general. Hence the form of $\mathcal{I}(\theta)$ cannot be determined. In pure sampling models, where all the variables are endogenous, this problem does not arise. However, the computation of $\mathcal{I}(\theta)$ may be complex. Thus one might also benefit by replacing $\mathcal{I}(\hat{\theta}_n)$ by either one of the following two matrices

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(y_i;\hat{\theta}_n)}{\partial\theta\,\partial\theta'}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(y_i;\hat{\theta}_n)}{\partial\theta}\frac{\partial \log f(y_i;\hat{\theta}_n)}{\partial\theta'}.$$

Remark 17.1: It is important to stress that the Wald test requires the maximization of the unconstrained likelihood function. On the other hand, the Wald test does not require the maximization of the constrained likelihood function.

Remark 17.2: A special case, which arises frequently in practice, is when $g(\theta) = \theta_1 - \theta_1^0$, where θ_1 is a subvector consisting of r elements of θ and θ_1^0 is a given r-dimensional vector. Without loss of generality, we can assume that these r elements are the first elements of θ. In this case, the test statistic is of the form

$$\xi_n^W = n(\hat{\theta}_{1n} - \theta_1^0)'(\mathcal{I}^{11}(\hat{\theta}_n))^{-1}(\hat{\theta}_{1n} - \theta_1^0), \tag{17.3}$$

where $\hat{\theta}_{1n}$ is the unconstrained maximum likelihood estimator of θ_1 and $\mathcal{I}^{11}(\theta)$ is the $r \times r$ northwest block of the inverse of the matrix $\mathcal{I}(\theta)$. If the information matrix $\mathcal{I}(\theta)$ is partitioned as

$$\mathcal{I}(\theta) = \begin{pmatrix} \mathcal{I}_{11}(\theta) & \mathcal{I}_{12}(\theta) \\ \mathcal{I}_{21}(\theta) & \mathcal{I}_{22}(\theta) \end{pmatrix}.$$

then $\mathcal{I}^{11}$ is equal to $(\mathcal{I}_{11} - \mathcal{I}_{12}\,\mathcal{I}_{22}^{-1}\mathcal{I}_{21})^{-1}$.

17.2.2 The Score or Lagrange Multiplier Test

Let $\hat{\theta}_n^0$ denote the maximum likelihood estimator of θ under the constraint $g(\theta) = 0$. The Lagrangian function associated with the constrained optimization problem is $L_n(\theta) + g'(\theta)\lambda$.

The first-order conditions are

$$\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} + \frac{\partial g'(\hat{\theta}_n^0)}{\partial \theta}\hat{\lambda}_n = 0, \tag{17.4}$$

where $\hat{\lambda}_n$ denotes the vector of Lagrange multipliers.

Property 17.2: *Under the assumptions of Property 7.17, the test defined by the critical region*

$$\left\{\xi_n^S \geq \chi^2_{1-\alpha}(r)\right\},$$

where

$$\begin{aligned}\xi_n^S &= \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\hat{\theta}_n^0)^{-1}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \\ &= \frac{1}{n}\hat{\lambda}_n'\frac{\partial g(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\hat{\theta}_n^0)^{-1}\frac{\partial g'(\hat{\theta}_n^0)}{\partial \theta}\hat{\lambda}_n,\end{aligned}$$

has asymptotic level α. This test is called the score or Lagrange multiplier test. In addition, under H_0, the difference $\xi_n^W - \xi_n^S$ converges in probability to zero as $n \to \infty$.

PROOF: As before, we use the notation # to indicate an equality that holds up to an additive term that converges to zero in probability.

a) SOME TAYLOR EXPANSIONS:

We consider some Taylor expansions around the true value θ_0. We have

$$\sqrt{n}g(\hat{\theta}_n) \# \frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0) \tag{17.5}$$

and

$$\sqrt{n}g(\hat{\theta}_n^0) \# \frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0). \tag{17.6}$$

Since $\hat{\theta}_n^0$ is the constrained estimator, we have $g(\hat{\theta}_n^0) = 0$. Thus, taking the difference of the preceding expressions, we obtain

$$\sqrt{n}g(\hat{\theta}_n) \# \frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0). \tag{17.7}$$

We now show that $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$ is asymptotically a linear function of $(1/\sqrt{n})\partial L_n(\hat{\theta}_n^0)/\partial\theta$. We have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n}{\partial \theta}(\hat{\theta}_n^0) \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \tag{17.8}$$

and

$$0 = \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{17.9}$$

Taking the difference, we obtain

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0). \tag{17.10}$$

Hence

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \# \mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}. \tag{17.11}$$

From (17.7) and (17.11) it follows that

$$\sqrt{n}g(\hat{\theta}_n) \# \frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta_0}.$$

Then, using (17.4), we obtain

$$\begin{aligned}\sqrt{n}g(\hat{\theta}_n) \quad &\# \quad -\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial g'(\hat{\theta}_n^0)}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}} \\ &\# \quad -\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}},\end{aligned}$$

i.e.

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \# -\left(\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}\sqrt{n}g(\hat{\theta}_n). \tag{17.12}$$

b) VARIOUS FORMS FOR ξ_n^S:
From (17.2) and (17.12) it follows that

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{d} N\left(0, \left(\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}\right)$$

as $n \to \infty$ under the null hypothesis H_0. The corresponding quadratic form is

$$\xi_n^S = \frac{1}{n}\hat{\lambda}_n'\frac{\partial g(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\hat{\theta}_n^0)^{-1}\frac{\partial g'(\hat{\theta}_n^0)}{\partial \theta}\hat{\lambda}_n. \tag{17.13}$$

Alternatively, using (17.4), another form for ξ_n^S is

$$\xi_n^S = \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\hat{\theta}_n^0)^{-1}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}. \tag{17.14}$$

c) ASYMPTOTIC EQUIVALENCE BETWEEN THE SCORE TEST AND THE WALD TEST:

From (17.12) we obtain

$$\xi_n^S \# \xi_n^W = ng'(\hat{\theta}_n) \left(\frac{\partial g(\hat{\theta}_n)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n)^{-1} \frac{\partial g'(\hat{\theta}_n)}{\partial \theta} \right)^{-1} g(\hat{\theta}_n).$$

under the null hypothesis H_0.

In particular, under H_0, the test statistic ξ_n^S has the same asymptotic distribution as ξ_n^W, namely a $\chi^2(r)$. Thus the test with critical region

$$W = \{\xi_n^S \geq \chi^2_{1-\alpha}(r)\}.$$

has asymptotic level α. □

Remark 17.3: The Lagrange multiplier test is consistent, in general. For, under the alternative hypothesis H_1, the estimator $\hat{\theta}_n^0$ converges to the pseudo true value θ_∞. Since $g(\hat{\theta}_n^0) = 0$, $\forall n$, we have $g(\theta_\infty) = 0$. Thus $\theta_\infty \neq \theta_0$ because $g(\theta_0) \neq 0$ under H_1. Moreover, $(1/n)\partial L_n(\hat{\theta}_n^0)/\partial\theta$ converges to $E_x E_0 \partial \log f(Y \mid X; \theta_\infty)/\partial\theta$. Because θ_0 is asymptotically identified and $\theta_\infty \neq \theta_0$, then θ_∞ cannot maximize $E_x E_0 \log f(Y \mid X; \theta)$ given that the latter is uniquely maximized at θ_0. Thus we have, in general

$$E_x E_0 \frac{\partial \log f(Y \mid X; \theta_\infty)}{\partial \theta} \neq 0.$$

This ensures the consistency of the Lagrange multiplier test in general.

Remark 17.4: Our previous remark on the estimation of $\mathcal{I}(\theta_0)$ still applies. Specifically, we can replace $\mathcal{I}(\hat{\theta}_n^0)$ by

$$-\frac{1}{n} \frac{\partial^2 \log L_n(\hat{\theta}_n^0)}{\partial\theta \, \partial\theta'}$$

or by

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(y_i \mid x_i; \hat{\theta}_n^0)}{\partial\theta} \frac{\partial \log f(y_i \mid x_i; \hat{\theta}_n^0)}{\partial\theta'}.$$

When the latter estimator of $\mathcal{I}(\theta_0)$ is used, the statistic ξ_n^S becomes

$$\xi_n^S = \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n^0)}{\partial \theta}\frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n^0)}{\partial \theta'}\right)^{-1}$$
$$\times \frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}$$
$$= \sum_{i=1}^n \frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n^0)}{\partial \theta'}\left(\sum_{i=1}^n \frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n^0)}{\partial \theta}\frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n^0)}{\partial \theta'}\right)^{-1}$$
$$\times \sum_{i=1}^n \frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n^0)}{\partial \theta}.$$

This allows us to state the next property.

Property 17.3: *Let e denote the vector of $\mathbb{R}^n$ whose components are all equal to 1. Let $\mathbf{F}$ denote the $n \times p$ matrix whose (i,j)-th element is $\partial \log f(y_i \mid x_i;\hat{\theta}_n^0)/\partial \theta_j$, where θ_j is the j-th component of θ. Then*

$$\xi_n^S = e'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'e = nR^2,$$

where R^2 is the coefficient of determination of the linear regression of e on the columns of $\mathbf{F}$ without the constant term.

Remark 17.5: Because we have replaced θ_0 by its estimator $\hat{\theta}_n^0$ under H_0 in the expression for the asymptotic variance covariance matrix of $\hat{\lambda}_n/\sqrt{n}$, the computation of ξ_n^S only requires the maximization of the likelihood function under the null hypothesis H_0.

Remark 17.6: When $g(\theta) = \theta_1 - \theta_1^0$, where θ_1 is the vector whose components are the first r elements of θ and θ_1^0 is a given r-dimensional vector, the statistic ξ_n^S becomes

$$\xi_n^S = \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta_1'}\mathcal{I}^{11}(\hat{\theta}_n^0)\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta_1}, \tag{17.15}$$

where $\mathcal{I}^{11}(\theta)$ is the $r \times r$ northwest submatrix of $\mathcal{I}^{-1}(\theta)$.

17.2.3 The Likelihood Ratio Test

This test is based on the comparison of the maximal values of the constrained and unconstrained likelihood function.

Property 17.4: *Under the assumptions of Property 7.17, the test defined by the critical region*

$$\{\xi_n^R \geq \chi^2_{1-\alpha}(r)\},$$

where

$$\xi_n^R = 2(L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0)),$$

has asymptotic level α and is consistent. This test is called the likelihood ratio test. In addition, under H_0, the differences $\xi_n^R - \xi_n^S$ and $\xi_n^R - \xi_n^W$ converge in probability to zero as $n \to \infty$.

PROOF: We first establish some useful asymptotic results.

a) SOME TAYLOR EXPANSIONS:

We consider the Taylor expansions of $L_n(\hat{\theta}_n)$ and $L_n(\hat{\theta}_n^0)$ around $\theta_0 \in \Theta$. Under H_0, these expansions give

$$L_n(\hat{\theta}_n) \quad \# \quad L_n(\theta_0) + \frac{\partial L_n(\theta_0)}{\partial \theta'}(\hat{\theta}_n - \theta_0) - \frac{n}{2}(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \tag{17.16}$$

and

$$L_n(\hat{\theta}_n^0) \# L_n(\theta_0) + \frac{\partial L_n(\theta_0)}{\partial \theta'}(\hat{\theta}_n^0 - \theta_0) - \frac{n}{2}(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0). \tag{17.17}$$

Taking the difference of these equalities, we obtain

$$\begin{aligned} \xi_n^R \quad \# \quad & 2\frac{\partial L_n(\theta_0)}{\partial \theta'}(\hat{\theta}_n - \hat{\theta}_n^0) - n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \\ & + n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0). \end{aligned} \tag{17.18}$$

But from (17.9), we have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{17.19}$$

Substituting this expression into (17.18), we obtain

$$\begin{aligned} \xi_n^R \quad \# \quad & 2n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0) - n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \\ & + n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0), \end{aligned} \tag{17.20}$$

which gives

$$\begin{aligned}\xi_n^R \quad \# \quad & 2n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0) - n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \\ & + n(\hat{\theta}_n^0 - \hat{\theta}_n + \hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \hat{\theta}_n + \hat{\theta}_n - \theta_0). \qquad (17.21)\end{aligned}$$

Then, developing the last term and cancelling terms gives

$$\xi_n^R \# n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0). \qquad (17.22)$$

b) ASYMPTOTIC EQUIVALENCES:
From (17.10) and (17.22), we obtain

$$\xi_n^R \# \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}, \qquad (17.23)$$

which shows that ξ_n^R is asymptotically equivalent to ξ_n^S. Hence, from Property 17.2, we also have

$$\xi_n^R \# \xi_n^W.$$

That is, ξ_n^R has the same asymptotic distribution as ξ_n^S and ξ_n^W under H_0.

c) CONSISTENCY OF THE LIKELIHOOD RATIO TEST:
To establish the consistency of the likelihood ratio test, we note that

$$\frac{L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0)}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(\log f(y_i \mid x_i; \hat{\theta}_n) - \log f(y_i \mid x_i; \hat{\theta}_n^0)\right)$$

Under H_1, this quantity converges to

$$E_x E_0 \left(\log f(Y \mid X; \theta_0) - \log f(Y \mid X; \theta_\infty)\right).$$

From Kullback inequality and the asymptotic identifiability of θ_0, it follows that this quantity is strictly positive. Hence ξ_n^R converges to $+\infty$, i.e., the test is consistent. □

The likelihood ratio test requires the constrained and unconstrained maximization of the likelihood function. On the other hand, the test statistic ξ_n^R does not involve the information matrix $\mathcal{I}(\theta_0)$ and hence does not require estimates of such a matrix.

17.2.4 Equivalent Tests and Independence Property

The three test statistics ξ_n^W, ξ_n^S, and ξ_n^R introduced in the previous sections are asymptotically equivalent under H_0 to various other statistics that may be theoretically or practically advantageous.

A first pair of asymptotically equivalent test statistics results immediately from the proof of Property 17.4.

Property 17.5: *The tests with critical regions*

$$\left\{n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{I}(\hat{\theta}_n^0)(\hat{\theta}_n - \hat{\theta}_n^0) \geq \chi^2_{1-\alpha}(r)\right\}$$

and

$$\left\{n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - \hat{\theta}_n^0) \geq \chi^2_{1-\alpha}(r)\right\}$$

are asymptotically equivalent to the Wald, score, and likelihood ratio tests under H_0.

PROOF: This is an immediate consequence of (17.22). □

These asymptotically equivalent test statistics will be useful when proving Property 17.7, which deals with independence.

Another asymptotically equivalent test statistic is a modified version of the likelihood ratio test statistic. A convenient feature of this new statistic is that it depends on $\hat{\theta}_n$ but not on $\hat{\theta}_n^0$.

Property 17.6: *The test with critical region*

$$\left\{2(L_n(\hat{\theta}_n) - L_n(\tilde{\theta}_n^0)) \geq \chi^2_{1-\alpha}(r)\right\},$$

where

$$\tilde{\theta}_n^0 = \hat{\theta}_n - \mathcal{I}(\hat{\theta}_n)^{-1}\frac{\partial g'(\hat{\theta}_n)}{\partial \theta}\left(\frac{\partial g(\hat{\theta}_n)}{\partial \theta'}\mathcal{I}(\hat{\theta}_n)^{-1}\frac{\partial g'(\hat{\theta}_n)}{\partial \theta}\right)^{-1} g(\hat{\theta}_n).$$

is asymptotically equivalent to the Wald, score, and likelihood ratio tests under H_0.

PROOF: From Property 10.9 we know that

$$\sqrt{n}(\hat{\theta}_n^0 - \tilde{\theta}_n^0) \# 0 \tag{17.24}$$

under H_0.

On the other hand, the difference between the second-order Taylor expansions of $L_n(\hat{\theta}_n^0)$ and $L_n(\tilde{\theta}_n^0)$ around θ_0 gives

$$L_n(\hat{\theta}_n^0) - L_n(\tilde{\theta}_n^0) \quad \# \quad \frac{\partial L_n(\theta_0)}{\partial \theta'}(\hat{\theta}_n^0 - \tilde{\theta}_n^0) - \frac{n}{2}(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) + \frac{n}{2}(\tilde{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\tilde{\theta}_n^0 - \theta_0).$$

Hence

$$L_n(\hat{\theta}_n^0) - L_n(\tilde{\theta}_n^0) \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^0 - \tilde{\theta}_n^0) - \frac{n}{2}(\hat{\theta}_n^0 - \tilde{\theta}_n^0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \tilde{\theta}_n^0) - n(\hat{\theta}_n^0 - \tilde{\theta}_n^0)'\mathcal{I}(\theta_0)(\tilde{\theta}_n^0 - \theta_0).$$

Using (17.24) and the convergence in distribution of $(1/\sqrt{n})\partial L_n(\theta_0)/\partial\theta$ and $\sqrt{n}(\tilde{\theta}_n^0 - \theta_0)$, we obtain $L_n(\hat{\theta}_n^0) - L_n(\tilde{\theta}_n^0) \# 0$. The desired result follows from Property 17.4. □

When the "approximate" constrained estimator $\tilde{\theta}_n^0$ is available, Property 17.6 allows to construct a testing procedure that is asymptotically equivalent to the classical testing procedures without computing the constrained maximum likelihood estimator $\hat{\theta}_n^0$.

Remark 17.7: A similar substitution of $\hat{\theta}_n^0$ by $\tilde{\theta}_n^0$ in the expression of the score test statistic ξ_n^S might be contemplated. However, because computations are more involved than in the preceding case, such a substitution is of limited interest.

Remark 17.8: As for the Wald test, the test defined in Property 17.6 only requires the maximization of the unconstrained likelihood function. In terms of computational complexity, these two procedures are therefore comparable.

Lastly, we give an independence property of the three classical tests. This property will be useful when considering the problem of sequential testing in Chapter 19.

Property 17.7: *The test statistics ξ_n^W, ξ_n^S, ξ_n^R, and their asymptotically equivalent test statistics given in Properties 17.5 and 17.6 are asymptotically independent of $\hat{\theta}_n^0$ under H_0.*

PROOF: From equation (17.22) and Property 17.5 we know that all the preceding test statistics are asymptotically equivalent to

$$n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0)$$

under H_0. Thus it suffices to establish the asymptotic independence of $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$ and $\sqrt{n}(\hat{\theta}_n^0 - \theta_0)$.

Equation (17.10) can be written

$$\mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) - \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta},$$

or alternatively, using equation (17.4)

$$\mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + \frac{\partial g'(\hat{\theta}_n^0)}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}}.$$

Then, using equation (17.12), we obtain

$$\begin{aligned}\mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \quad \# \quad & \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) \\ & -\frac{\partial g'(\theta_0)}{\partial \theta}\left(\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}\sqrt{n}g(\hat{\theta}_n).\end{aligned}$$

Now, using equation (17.5), we obtain

$$\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \# \mathbf{M}'(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0),$$

where

$$\mathbf{M}(\theta_0) \quad = \quad \mathbf{I} - \frac{\partial g'(\theta_0)}{\partial \theta}\left(\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}.$$

Hence

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \# (\mathbf{I} - \mathbf{M}'(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{17.25}$$

Therefore the pair of vectors $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$ and $\sqrt{n}(\hat{\theta}_n^0 - \theta_0)$ is jointly asymptotically normally distributed with asymptotic variance covariance matrix

$$(\mathbf{I} - \mathbf{M}'(\theta_0))\mathcal{I}(\theta_0)^{-1}\mathbf{M}(\theta_0) = 0.$$

This is because the matrix $\mathbf{M}(\theta_0)$ is the orthogonal projection on to the subspace that is orthogonal to the subspace generated by the columns of $\partial g'(\theta_0)/\partial\theta$ with respect to the scalar product defined by $\mathcal{I}(\theta_0)^{-1}$. The desired independence property follows. □

17.3 Examples of Tests of a Null Hypothesis in Impicit Form

17.3.1 The Linear Model With Known Variance

We consider the linear model

$$Y = \mathbf{X}\theta + u,$$

where Y is $n \times 1$, $\mathbf{X}$ is $n \times K$, θ is $K \times 1$, and u is $n \times 1$. The distribution of u given $\mathbf{X}$ is the multivariate normal distribution $N(0, \sigma_0^2 \mathbf{I})$, where σ_0^2 is known and $\mathbf{X}$ is of rank K.

We want to test the linear hypothesis defined by

$$H_0 : \mathbf{G}\theta = g,$$

where $\mathbf{G}$ is of rank r and g a given $r \times 1$ vector.

The log-likelihood function (conditionally on $\mathbf{X}$) is

$$L_n(\theta) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}(Y - \mathbf{X}\theta)'(Y - \mathbf{X}\theta).$$

The unconstrained maximum likelihood estimator of θ is the ordinary least squares estimator

$$\hat{\theta}_n = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y. \tag{17.26}$$

The constrained maximum likelihood estimator of θ under H_0 is the constrained least squares estimator

$$\hat{\theta}_n^0 = \hat{\theta}_n + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'\left(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'\right)^{-1}(g - \mathbf{G}\hat{\theta}_n) \tag{17.27}$$

(see (equation 10.11)).

The vector of Lagrange multipliers corresponding to the constraints $\mathbf{G}\theta = g$ is

$$\hat{\lambda}_n = -\frac{2}{\sigma_0^2}(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}(g - \mathbf{G}\hat{\theta}_n). \tag{17.28}$$

a) Wald Test

The Wald test is based on $\mathbf{G}\hat{\theta}_n - g$. Under H_0, we have

$$\mathbf{G}\hat{\theta}_n - g \sim N(0, \sigma_0^2\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}').$$

The Wald statistic uses the corresponding quadratic form. This gives

$$\xi_n^W = \frac{1}{\sigma_0^2}(\mathbf{G}\hat{\theta}_n - g)'(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}(\mathbf{G}\hat{\theta}_n - g) \sim \chi^2(r). \quad (17.29)$$

Note that the statistic ξ_n^W thus obtained is numerically equal to the statistic defined in Property 17.1 provided $I(\hat{\theta}_n)$ is replaced by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(y_i \mid x_i; \hat{\theta}_n)}{\partial\theta\ \partial\theta'} = \frac{1}{n\sigma_0^2}\mathbf{X}'\mathbf{X}.$$

In particular, *the $\chi^2(r)$ distribution of the Wald statistic ξ_n^W holds not only approximately as $n \to \infty$ but also exactly in finite samples.*

b) Score or Lagrange Multiplier Test

Similarly, under H_0, the Lagrange multiplier $\hat{\lambda}_n$ is distributed as

$$N\left(0, \frac{4}{\sigma_0^2}(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}\right).$$

Thus the score statistic is

$$\xi_n^S = \frac{\sigma_0^2}{4}\hat{\lambda}_n\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'\hat{\lambda}_n. \quad (17.30)$$

It is easy to verify that the two test statistics ξ_n^S and ξ_n^W are numerically equal.

c) Likelihood Ratio Test

The difference between the constrained and the unconstrained values of the likelihood function is

$$L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0) = -\frac{1}{2\sigma_0^2}\left(\|Y - \mathbf{X}\hat{\theta}_n\|^2 - \|Y - \mathbf{X}\hat{\theta}_n^0\|^2\right),$$

which gives, using Pythagora's Theorem

$$L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0) = \frac{1}{2\sigma_0^2}\|\mathbf{X}\hat{\theta}_n^0 - \mathbf{X}\hat{\theta}_n\|^2.$$

Hence

$$\xi_n^R = \frac{1}{\sigma_0^2}\|\mathbf{X}\hat{\theta}_n^0 - \mathbf{X}\hat{\theta}_n\|^2. \quad (17.31)$$

Moreover, equation (17.27) implies

$$\begin{aligned}\xi_n^R &= \frac{1}{\sigma_0^2}\|\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}(g-\mathbf{G}\hat{\theta}_n)\|^2 \\ &= \frac{1}{\sigma_0^2}(g-\mathbf{G}\hat{\theta}_n)'(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}(g-\mathbf{G}\hat{\theta}_n) \\ &= \xi_n^W.\end{aligned}$$

To summarize, the three test statistics ξ_n^W, ξ_n^S, and ξ_n^R, which are asymptotically equivalent, in general, are here *numerically equal in finite samples.* Moreover, the sizes of the corresponding tests are exactly equal to α.

d) Independence Between the Constrained Estimator $\hat{\theta}_n^0$ and the Classical Test Statistics

The classical test statistics depend on the observations through $\mathbf{G}\hat{\theta}_n - g$. On the other hand, the constrained estimator $\hat{\theta}_n^0$ is equal to

$$\hat{\theta}_n^0 = \left(\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}\mathbf{G}\right)\hat{\theta}_n + cst.$$

Hence the statistics $\mathbf{G}\hat{\theta}_n - g$ and $\hat{\theta}_n^0$ are jointly normally distributed. They are mutually independent if their covariance matrix is equal to zero. We have

$$\begin{aligned}\mathrm{Cov}(\hat{\theta}_n^0, \mathbf{G}\hat{\theta}_n - g) &= \sigma_0^2\left(\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}\mathbf{G}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}' \\ &= 0\end{aligned}$$

as desired. In this case, the mutual independence between $\hat{\theta}_n^0$ and the classical test statistics holds also exactly in finite samples. Thus such a result strengthens Property 17.7.

17.3.2 The Linear Model With Unknown Variance

To simplify the exposition, we keep the notation θ for the vector of coefficients associated with the explanatory variables although there is an additional parameter, which is σ^2.

The unconstrained and constrained estimators $\hat{\theta}_n$ and $\hat{\theta}_n^0$ are as before. The corresponding estimators of σ^2 are

$$\hat{\sigma}_n^2 = \frac{1}{n}\|Y - \mathbf{X}\hat{\theta}_n\|^2$$

and

$$\hat{\sigma}_0^2 = \frac{1}{n}\|Y - \mathbf{X}\hat{\theta}_n^0\|^2.$$

respectively.

a) Wald Test

The Wald test statistic becomes

$$\xi_n^W = \frac{1}{\hat{\sigma}_n^2}(\mathbf{G}\hat{\theta}_n - g)'(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}(\mathbf{G}\hat{\theta}_n - g).$$

From Pythagora's Theorem, we obtain

$$\begin{aligned}\hat{\sigma}_0^2 - \hat{\sigma}_n^2 &= \frac{1}{n}\|\mathbf{X}\hat{\theta}_n^0 - \mathbf{X}\hat{\theta}_n\|^2 \\ &= \frac{1}{n}(\mathbf{G}\hat{\theta}_n - g)'(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}')^{-1}(\mathbf{G}\hat{\theta}_n - g).\end{aligned}$$

Thus

$$\xi_n^W = n\frac{\hat{\sigma}_0^2 - \hat{\sigma}_n^2}{\hat{\sigma}_n^2} = \frac{nr}{n-K}F, \tag{17.32}$$

where F is the Fisher statistic defined in Chapter 16.

b) Score Test

The score test statistic becomes

$$\xi_n^S = n\frac{\hat{\sigma}_0^2}{4}\hat{\lambda}_n'\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'\hat{\lambda}_n, \tag{17.33}$$

where

$$\hat{\lambda}_n = \frac{-2}{\hat{\sigma}_0^2}\left(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'\right)^{-1}\left(g - \mathbf{G}\hat{\theta}_n\right).$$

The score test statistic can also be expressed in terms of the Fisher statistic. We have

$$\xi_n^S = \frac{1}{\hat{\sigma}_0^2}\left(\mathbf{G}\hat{\theta}_n - g\right)'\left(\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}'\right)^{-1}\left(\mathbf{G}\hat{\theta}_n - g\right).$$

This gives

$$\xi_n^S = n\frac{\hat{\sigma}_0^2 - \hat{\sigma}_n^2}{\hat{\sigma}_0^2} = \frac{n}{1 + (n-K)/rF}. \tag{17.34}$$

c) Likelihood Ratio Test

We have

$$L_n(\hat{\theta}_n, \hat{\sigma}_n^2) - L_n(\hat{\theta}_n^0, \hat{\sigma}_0^2) = \frac{n}{2} \log \frac{\hat{\sigma}_0^2}{\hat{\sigma}_n^2}.$$

Thus

$$\xi_n^R = n \log \frac{\hat{\sigma}_0^2}{\hat{\sigma}_n^2} = n \log(1 + \frac{rF}{n-K}). \tag{17.35}$$

d) Finite Sample Distributions

From the general theory, the three test statistics ξ_n^W, ξ_n^S, and ξ_n^R are asymptotically distributed as $\chi^2(r)$ under H_0. For the linear model with unknown variance, the preceding results show that the quantiles of the distributions of these statistics can be computed *exactly* in finite samples. This is because all three statistics are increasing functions of the Fisher statistic, which follows a Fisher distribution with r and $n-K$ degrees of freedom.

e) Comparing of the Classical Test Statistics

Using the inequalities

$$\frac{x}{1+x} \leq \log(1+x) \leq x, \quad \forall\, x > -1,$$

we obtain with $x = (\hat{\sigma}_0^2 - \hat{\sigma}_n^2)/\hat{\sigma}_n^2$

$$\xi_n^S \leq \xi_n^R \leq \xi_n^W. \tag{17.36}$$

Note also that $x = (\hat{\sigma}_0^2 - \hat{\sigma}_n^2)/\hat{\sigma}_n^2$ converges asymptotically to zero under H_0. Because the three functions $x/(1+x)$, $\log(1+x)$, and x are equivalent to x in a neighborhood of zero, it follows at once that our three test statistics are asymptotically equivalent in the present example.

f) Interpretation in Terms of the Coefficient of Determination

We note that the score test statistic ξ_n^S can be written as

$$\xi_n^S = n \frac{\|\hat{u}_0\|^2 - \|\hat{u}\|^2}{\|\hat{u}_0\|^2},$$

where $\hat{u}_0 = Y - \mathbf{X}\hat{\theta}_n^0$ and $\hat{u} = Y - \mathbf{X}\hat{\theta}_n$ are the vectors of residuals of the constrained and unconstrained regressions respectively.

Note also that an alternative expression for ξ_n^S is

$$\xi_n^S = \frac{1}{\hat{\sigma}_0^2}\|\mathbf{P}\hat{u}_0\|^2 = \frac{1}{\hat{\sigma}_0^2}\hat{u}_0'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{u}_0, \tag{17.37}$$

where $\mathbf{P}$ is the orthogonal projection matrix on the subspace generated by the columns of $\mathbf{X}$.

It follows that

$$\xi_n^S = nR^2, \tag{17.38}$$

where R^2 is the (noncentered) coefficient of determination of the linear regression of $\hat{u}_0$ on the columns of $\mathbf{X}$.

17.3.3 The Nonlinear Gaussian Regression Model

We consider the nonlinear Gaussian regression model

$$Y_i = m(x_i, \theta) + u_i, \quad i = 1, \ldots, n,$$

where x_i is a K-dimensional vector, θ is a p-dimensional vector of parameters, and $m(x_i, \theta)$ is a differentiable function in θ.

Conditionally upon x_i, $i = 1, \ldots, n$, the joint distribution of the vector u whose i-th component is u_i, $i = 1, \ldots, n$, is the multivariate normal distribution $N(0, \sigma^2\mathbf{I})$, where σ is an unknown parameter. As in the preceding subsection, the vector θ does not include all the parameters of the model.

We want to test the null hypothesis H_0: $g(\theta) = 0$ against the alternative hypothesis H_1: $g(\theta) \neq 0$.

a) Fisher Information Matrix

We have

$$\begin{aligned} L_n(\theta, \sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}u'u \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - m(x_i, \theta))^2. \end{aligned}$$

Hence

$$\begin{aligned} \frac{\partial L_n(\theta, \sigma^2)}{\partial \theta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{\partial m(x_i, \theta)}{\partial \theta}(y_i - m(x_i, \theta)) \\ &= \frac{1}{\sigma^2}\mathbf{M}'u, \end{aligned}$$

where $\mathbf{M}$ is the $n \times p$ matrix whose generic term is $\partial m(x_i, \theta)/\partial \theta_j$. The matrix $\mathbf{M}$ is assumed to be of rank p.

For the second-order derivatives we obtain

$$\begin{aligned} \frac{\partial^2 L_n}{\partial \theta \, \partial \theta'} &= -\frac{1}{\sigma^2}\mathrm{M}'\mathrm{M}, \\ \frac{\partial^2 L_n}{\partial \theta \, \partial \sigma^2} &= -\frac{1}{\sigma^4}\mathrm{M}'u. \end{aligned}$$

Taking expectations of minus those derivatives, we obtain

$$\begin{aligned} -E_\theta \frac{\partial^2 L_n}{\partial \theta \, \partial \theta'} &= \frac{1}{\sigma^2}\mathrm{M}'\mathrm{M}, \\ -E_\theta \frac{\partial^2 L_n}{\partial \theta \, \partial \sigma^2} &= 0. \end{aligned}$$

Therefore the Fisher information matrix is block diagonal. Consequently, the term $-E_\theta(\partial^2 L_n/\partial\sigma^2 \, \partial\sigma^2)$ does not appear in the classical test statistics, which are derived below.

b) Wald Test

We have

$$\xi_n^W = \frac{1}{\hat{\sigma}_n^2} g'(\hat{\theta}_n) \left(\frac{\partial g(\hat{\theta}_n)}{\partial \theta'} (\hat{\mathrm{M}}'\hat{\mathrm{M}})^{-1} \frac{\partial g'(\hat{\theta}_n)}{\partial \theta} \right)^{-1} g(\hat{\theta}_n), \tag{17.39}$$

where $\hat{\theta}_n$ is the unconstrained maximum likelihood estimator of θ, i.e., the nonlinear least squares estimator, $\hat{\sigma}_n^2 = (\hat{u}'\hat{u})/n$ is the estimator of σ^2 with $\hat{u}$ being the vector of unconstrained residuals $\hat{u}_i = Y_i - m(x_i, \hat{\theta}_n)$, and $\hat{\mathrm{M}}$ is the matrix with generic term $\partial m(x_i, \hat{\theta}_n)/\partial \theta_j$.

c) Score Test

The score test statistic can be written as

$$\xi_n^S = \frac{1}{\hat{\sigma}_0^2} \hat{u}_0' \hat{\mathrm{M}}_0 (\hat{\mathrm{M}}_0' \hat{\mathrm{M}}_0)^{-1} \hat{\mathrm{M}}_0' \hat{u}_0, \tag{17.40}$$

where $\hat{u}_0$ is the vector of constrained residuals $\hat{u}_{i0} = \hat{Y}_i - m(x_i, \hat{\theta}_n^0)$, $\hat{\mathrm{M}}_0$ is the matrix with generic term $\partial m(x_i, \hat{\theta}_n^0)/\partial \theta_j$, and $\hat{\sigma}_0^2$ is the constrained estimator $\hat{u}_0'\hat{u}_0/n$ of the variance σ^2.

The score test statistic can also be written as

$$\xi_n^S = nR^2, \tag{17.41}$$

where R^2 is the (noncentered) coefficient of determination of the regression of $\hat{u}^0$ on the columns of the matrix $\hat{M}_0$.

Expressions (17.40) and (17.41) for the score test statistic are analogous to those obtained for the linear model with unknown variance (see equations (17.37) and (17.38)), where $\mathbf{X}$ is replaced by $\hat{\mathbf{M}}_0$.

17.4 Wald, Score, and Likelihood Ratio Tests When the Null Hypothesis Is of the Form $\theta = h(\gamma)$ (Explicit Form)

We consider again the general parametric model of Section 17.2 and we are interested in asymptotic tests based on the conditional likelihood function. The null hypothesis is now of the form

$$\Theta_0 = \{\theta : \theta = h(\gamma), \gamma \in \Gamma\}, \text{ where } h(\gamma) = \begin{pmatrix} h_1(\gamma) \\ \vdots \\ h_p(\gamma) \end{pmatrix}. \tag{17.42}$$

Here γ is a parameter vector varying in $\Gamma \subset \mathbb{R}^q$ and h is a differentiable function from $\mathbb{R}^q$ to $\mathbb{R}^p$. The preceding form of the null hypothesis is said to be *explicit*.

17.4.1 A Wald-Type Test

In Section 17.2 we saw that, when the constraints defining the null hypothesis H_0 are written in implicit form, i.e., when H_0: $g(\theta) = 0$, the basic idea of the Wald test is to compare $g(\hat{\theta}_n)$ to the value of $g(\theta)$ under H_0, i.e., to zero. Following this idea when the constraints of the null hypothesis are in the explicit form $\theta = h(\gamma)$, it is natural to compare $\hat{\theta}_n$ to $\hat{\theta}_n^0 = h(\hat{\gamma}_n)$, where $\hat{\gamma}_n$ is the maximum likelihood estimator of γ.

Property 17.8: *Under the assumptions of Property 7.17, if*

$$\operatorname{rank}\left(\frac{\partial h'(\gamma)}{\partial \gamma} \mathcal{I}(h(\gamma)) \frac{\partial h(\gamma)}{\partial \gamma'}\right) = q,$$

then the test with critical region

$$\{\xi_n^W \geq \chi^2_{1-\alpha}(p-q)\},$$

where

$$\xi_n^W = n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - \hat{\theta}_n^0),$$

has asymptotic level α and is consistent.

PROOF: We shall first determine the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$ for a parameter value $\theta_0 = h(\gamma_0)$.

a) TAYLOR EXPANSION OF $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$:
Let $L_n^*(\gamma) = L_n(h(\gamma))$. Then, taking a Taylor expansion of the quantity $\partial L_n^*(\hat{\gamma}_n)/\partial\gamma$ around γ_0, we obtain

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\frac{\partial L_n^*(\hat{\gamma}_n)}{\partial\gamma} \\ &\# \frac{\partial h'(\gamma_0)}{\partial\gamma}\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta} \\ &\quad - \left(\frac{\partial h'(\gamma_0)}{\partial\gamma}\mathcal{I}(\theta_0)\frac{\partial h(\gamma_0)}{\partial\gamma'} - \sum_{i=1}^{p}\frac{\partial^2 h_i(\gamma_0)}{\partial\gamma\,\partial\gamma'}E_x E_0\frac{\partial \log f(Y \mid X;\theta_0)}{\partial\theta_i}\right) \\ &\quad \times\sqrt{n}(\hat{\gamma}_n - \gamma_0). \end{aligned}$$

Under the null hypothesis H_0, we have $E_x E_0(\partial \log f(Y \mid X;\theta_0)/\partial\theta) = 0$. Thus

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \# \left(\frac{\partial h'(\gamma_0)}{\partial\gamma}\mathcal{I}(\theta_0)\frac{\partial h(\gamma_0)}{\partial\gamma'}\right)^{-1}\frac{\partial h'(\gamma_0)}{\partial\gamma}\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta}.$$

b) EXPRESSING $\hat{\gamma}_n - \gamma_0$ AS A FUNCTION OF $\hat{\theta}_n - \theta_0$:
Maximum likelihood theory implies

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta} \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0).$$

Hence from a) we obtain

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \quad \# \quad \left(\frac{\partial h'(\gamma_0)}{\partial\gamma}\mathcal{I}(\theta_0)\frac{\partial h(\gamma_0)}{\partial\gamma'}\right)^{-1}\frac{\partial h'(\gamma_0)}{\partial\gamma}\mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{17.43}$$

c) EXPRESSING $\hat{\theta}_n^0 - \theta_0$ AS A FUNCTION OF $\hat{\theta}_n - \theta_0$:
We have

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n^0 - \theta_0) &= \sqrt{n}(h(\hat{\gamma}_n) - h(\gamma_0)) \\ &\# \frac{\partial h(\gamma_0)}{\partial \gamma'} \sqrt{n}(\hat{\gamma}_n - \gamma_0). \end{aligned}$$

Hence, using (17.43), we obtain

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n^0 - \theta_0) \quad \# \quad & \frac{\partial h(\gamma_0)}{\partial \gamma'} \left(\frac{\partial h'(\gamma_0)}{\partial \gamma} \mathcal{I}(\theta_0) \frac{\partial h(\gamma_0)}{\partial \gamma'} \right)^{-1} \\ & \times \frac{\partial h'(\gamma_0)}{\partial \gamma} \mathcal{I}(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0). \end{aligned} \tag{17.44}$$

This shows that $\hat{\theta}_n^0 - \theta_0$ and $\hat{\theta}_n - \theta_0$ are asymptotically linearly related, namely

$$\sqrt{n}(\hat{\theta}_n^0 - \theta) \# \mathbf{P}_h \sqrt{n}(\hat{\theta}_n - \theta_0), \tag{17.45}$$

where $\mathbf{P}_h$ denotes the matrix of the orthogonal projection on to the q-dimensional subspace of $\mathbb{R}^p$ generated by the columns of $\partial h(\gamma_0)/\partial \gamma'$ with respect to the scalar product defined by $\mathcal{I}(\theta_0)$.

d) EXPRESSING $\hat{\theta}_n - \hat{\theta}_n^0$ AS A FUNCTION OF $\hat{\theta}_n - \theta_0$:
We also have

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) &= \sqrt{n}(\hat{\theta}_n - \theta_0) - \sqrt{n}(\hat{\theta}_n^0 - \theta_0) \\ &\# \mathbf{M}_h \sqrt{n}(\hat{\theta}_n - \theta_0), \end{aligned} \tag{17.46}$$

where $\mathbf{M}_h = \mathbf{I} - \mathbf{P}_h$ is also an orthogonal projection matrix.

e) ASYMPTOTIC DISTRIBUTION OF $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$:
From above, it follows that the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$ is the multivariate normal distribution with zero mean and variance covariance matrix

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{M}_h \mathcal{I}^{-1}(\theta_0) \mathbf{M}_h' \\ &= \mathcal{I}(\theta_0)^{-1} \mathbf{M}_h' \\ &= \mathbf{M}_h \mathcal{I}(\theta_0)^{-1}. \end{aligned} \tag{17.47}$$

The latter two equalities follow from the fact that $\mathbf{M}_h$ is an orthogonal projection matrix with respect to the scalar product defined by $\mathcal{I}(\theta_0)$.

f) TEST STATISTIC:

The preceding results show that the matrix Σ is of rank $p-q$. Moreover, $\mathcal{I}(\theta_0)$ is a generalized inverse of Σ since

$$\Sigma\mathcal{I}(\theta_0)\Sigma = \mathrm{M}_h\mathcal{I}(\theta_0)^{-1}\mathcal{I}(\theta_0)\mathcal{I}(\theta_0)^{-1}\mathrm{M}_h' = \Sigma.$$

From Corollary B.5, it follows that the quadratic form

$$\xi_n^W \# n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0)$$

is asymptotically distributed as a χ^2 with $p-q$ degrees of freedom.

g) CONSISTENCY OF THE TEST:

Suppose that the true value θ_0 does not satisfy the null hypothesis H_0, i.e., suppose that θ_0 is no longer of the form $h(\gamma_0)$ for some γ_0. Then $\hat{\theta}_n$ converges to θ_0 while $\hat{\theta}_n^0$ converges to $\theta_\infty = h(\gamma_\infty) \neq \theta_0$. Thus ξ_n^W converges to $+\infty$. Hence the test is consistent. □

Note that Property 17.8 parallels Property 17.5, which was obtained when the constraints defining the null hypothesis were in implicit form. Unlike for implicit constraints, however, it is interesting to emphasize that the Wald statistic ξ_n^W depends on both the constrained and unconstrained maximum likelihood estimators.

17.4.2 Score Test

When the constraints are of the form $\theta = h(\gamma)$, Lagrange multipliers no longer exist. Nonetheless, it is possible to define a score test as the next property shows.

Property 17.9: *Under the assumptions of Property 17.8, the test defined by the critical region*

$$\{\xi_n^S \geq \chi^2_{1-\alpha}(p-q)\},$$

where

$$\xi_n^S = \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial\theta'}\mathcal{I}^{-1}(\hat{\theta}_n^0)\frac{\partial L_n(\hat{\theta}_n^0)}{\partial\theta},$$

has asymptotic level α. In addition, under H_0, the difference $\xi_n^S - \xi_n^W$ converges to zero in probability.

PROOF: We use the following Taylor expansion

$$0 = \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n)}{\partial\theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) \qquad (17.48)$$

and

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0). \tag{17.49}$$

Then we have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0), \tag{17.50}$$

i.e.

$$\frac{1}{\sqrt{n}}\mathcal{I}^{-1/2}(\theta_0)\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \mathcal{I}^{1/2}(\theta_0)\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0). \tag{17.51}$$

Taking norms and replacing θ_0 by $\hat{\theta}_n^0$ in the left-hand side and by $\hat{\theta}_n$ in the right-hand side, the asymptotic equivalence between ξ_n^S and ξ_n^W, i.e., $\xi_n^S \# \xi_n^W$, follows. In particular, the asymptotic distribution of ξ_n^S is the same as that of ξ_n^W under H_0. □

Suppose that θ_0 does not correspond to a parameter value satisfying the null hypothesis H_0. Then $\hat{\theta}_n^0$ converges to $\theta_\infty = h(\gamma_\infty) \neq \theta_0$. Hence $(1/n)(\partial L_n(\hat{\theta}_n^0)/\partial\theta)$ converges to $E_x E_0 \partial \log f(Y \mid X; \theta_\infty)/\partial\theta$. The latter quantity is typically nonzero when θ_0 is asymptotically identified. Therefore the above test is consistent, in general.

This score test is identical to the score test proposed in Property 17.2 and only requires the constrained maximization of the likelihood function. The remarks made about the estimation of $I(\theta_0)$ in Section 17.2.2 as well as the interpretation of ξ_n^S in terms of a R^2 still apply.

17.4.3 Likelihood Ratio Test

Property 17.10: *Under the assumptions of Property 17.8, the test defined by the critical region*

$$\left\{\xi_n^R \geq \chi^2_{1-\alpha}(p-q)\right\},$$

where

$$\xi_n^R = 2(L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0)),$$

has asymptotic level α and is consistent. In addition, under H_0, the differences $\xi_n^R - \xi_n^S$ and $\xi_n^R - \xi_n^W$ converge to zero in probability.

PROOF: The proof of Property 17.4, which establishes the asymptotic equivalence of ξ_n^R and $n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0)$ under H_0, does not depend on the form of the constraints. Thus this proof remains valid here. □

17.4.4 Equivalent Tests and Independence Property

In Chapter 10 , which deals with estimation under equality constraints (see especially Property 10.64), we saw that there exists a constrained estimator $\bar{\theta}_n^0$ satisfying $\sqrt{n}(\hat{\theta}_n^0 - \bar{\theta}_n^0) \# 0$ and which can be computed from the unconstrained estimator $\hat{\theta}_n$ only. This estimator is $\bar{\theta}_n^0 = h(\bar{\gamma}_n)$, where $\bar{\gamma}_n$ is obtained by minimizing with respect to γ

$$(\hat{\theta}_n - h(\gamma))'(\hat{V}_{\text{as}}(\hat{\theta}_n))^{-1}(\hat{\theta}_n - h(\gamma)),$$

i.e.

$$(\hat{\theta}_n - h(\gamma))'\mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - h(\gamma)),$$

for instance. As the next property shows, the constrained estimator $\bar{\theta}_n^0$ can replace the estimator $\hat{\theta}_n^0$ in the preceding test statistics.

Property 17.11: *Suppose that $\bar{\theta}_n^0$ replaces $\hat{\theta}_n^0$ in the test statistics ξ_n^W, ξ_n^S, and ξ_n^R of Properties 17.8, 17.9, and 17.10. Then the resulting test statistics are asymptotically equivalent to the original test statistics under H_0.*

PROOF: For ξ_n^R, the proof is identical to the proof of Property 17.6 because $\sqrt{n}(\hat{\theta}_n^0 - \bar{\theta}_n^0) \# 0$. For the other two statistics ξ_n^W and ξ_n^S, a similar proof can be given. □

Remark 17.9: Property 17.11 is more useful than its analog in the case of implicit constraints $g(\theta) = 0$. This is because, in the case of implicit constraints, the Wald test is based on a statistic using $\hat{\theta}_n$ only. This is no longer the case for the Wald-type test of Property 17.8 when H_0 is of the explicit form $\theta = h(\gamma)$. In particular, when the computation of $\hat{\theta}_n^0$ is difficult, then one can use instead $\bar{\theta}_n^0$, which is based on $\hat{\theta}_n$ only.

Property 17.11 will also be useful when studying sequential tests.

Remark 17.10: Property 17.11 and the form of ξ_n^W in Property 17.8 imply that a test statistic asymptotically equivalent to ξ_n^W, ξ_n^S, and ξ_n^R is

$$n(\hat{\theta}_n - h(\bar{\gamma}_n))'\mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - h(\bar{\gamma}_n)),$$

which is equal to n times the minimal value of the optimization problem

$$\min_{\gamma} (\hat{\theta}_n - h(\gamma))'\mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - h(\gamma))$$

that defines $\bar{\gamma}_n$.

Property 17.12: *Under H_0, the test statistics ξ_n^W, ξ_n^S, ξ_n^R, and their equivalent test statistics considered in Property 17.11 are asymptotically independent of $\hat{\theta}_n^0$.*

PROOF: Because

$$\xi_n^W \# \xi_n^R \# \xi_n^S \# \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta},$$

it suffices to show that $(1/\sqrt{n})(\partial L_n(\hat{\theta}_n^0)/\partial\theta)$ and $\sqrt{n}(\hat{\theta}_n^0 - \theta_0)$ are asymptotically independent. But from equations (17.50) and (17.51), we have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \mathcal{I}(\theta_0)\mathbf{M}_h\sqrt{n}(\hat{\theta}_n - \theta_0).$$

Now, equation (17.44) implies that the pair $(1/\sqrt{n})(\partial L_n(\hat{\theta}_n^0)/\partial\theta)$ and $\sqrt{n}(\hat{\theta}_n^0 - \theta_0)$ are jointly normally distributed with asymptotic variance covariance matrix

$$\mathcal{I}(\theta_0)\mathbf{M}_h\mathcal{I}(\theta_0)^{-1}\mathbf{P}_h' = 0.$$

This gives the desired result. □

17.5 Examples of Tests of a Null Hypothesis in Explicit Form: The Chi-Square Tests

17.5.1 Tests of Goodness-of-Fit to a Family of Distributions

Suppose that $Y_1, \ldots, Y_n$ are independent random variables with a common but unknown probability distribution P. Suppose also that P is a discrete distribution defined on a set of m elements. Without loss of generality, we can take this set to be $\{1, 2, \ldots, m\}$. Thus the distribution P is characterized by m positive real numbers p_j, $j = 1, \ldots, m$ satisfying, $\sum_{i=1}^m p_j = 1$.

We want to test the null hypothesis H_0, that P belongs to a given parametric family of distributions characterized by the vector $p(\gamma)$ whose components are $p_j(\gamma)$, $j = 1, \ldots, m$, with $p_j(\gamma) > 0$, $\sum_{i=1}^m p_j(\gamma) = 1$, and $\gamma \in \Gamma$, an open subset of $\mathbb{R}^q$. The functions $p_j(\gamma)$ are assumed to be differentiable and the $m \times q$ matrix $\partial p/\partial\gamma'$ is assumed to be of rank q.

The results of the preceding section cannot be applied directly. This is because θ cannot be the vector $p = (p_1, \ldots, p_m)'$ in view of the constraint $\sum_{i=1}^{m} p_j = 1$. Doing so would imply that the interior of the parameter space for θ is empty. This would invalidate the use of derivatives and the computation of Fisher information. We can, however, let $\theta = (p_1, \ldots, p_{m-1})'$. All the preceding results now apply to a parameter vector of dimension $m-1$.

The Fisher information matrix associated with θ thus defined is

$$\operatorname{diag}\left(\frac{1}{\theta}\right) + \frac{1}{1 - e'\theta} e\, e' = \operatorname{diag}\left(\frac{1}{\theta}\right) + \frac{1}{p_m} e\, e',$$

where $\operatorname{diag}(1/\theta)$ is the diagonal matrix of which the elements are $1/\theta_i$, $i = 1, \ldots, m-1$, and e is the $(m-1)$-dimensional vector whose components are all equal to one.

The Wald-type test statistic of Property 17.8 becomes

$$\begin{aligned} \xi_n^W &= n(\hat{\theta}_n - \hat{\theta}_n^0)' \left(\operatorname{diag}\left(\frac{1}{\hat{\theta}_n}\right) + \frac{1}{1 - e'\hat{\theta}_n} e\, e' \right) (\hat{\theta}_n - \hat{\theta}_n^0) \\ &= n \sum_{j=1}^{m-1} \frac{(\hat{p}_j - p_j(\hat{\gamma}_n))^2}{\hat{p}_j} + \frac{n}{\hat{p}_m} (e'(\hat{\theta}_n - \hat{\theta}_n^0))^2 \\ &= n \sum_{j=1}^{m} \frac{(\hat{p}_j - p_j(\hat{\gamma}_n))^2}{\hat{p}_j}, \end{aligned}$$

where $\hat{p}_j$ and $p_j(\hat{\gamma}_n)$ denote the unconstrained and constrained maximum likelihood estimators of p_j respectively. Hence the critical region of this Wald-type test is

$$\left\{ \xi_n^W \geq \chi^2_{1-\alpha}(m - 1 - q) \right\}.$$

Clearly, another Wald-type test, which is asymptotically equivalent under H_0, is obtained by considering the test statistic

$$\xi_n = n \sum_{j=1}^{m} \frac{(\hat{p}_j - p_j(\hat{\gamma}_n))^2}{\hat{p}_j(\hat{\gamma}_n)}.$$

Definition 17.3: *Consider the problem of testing the null hypothesis*

$$H_0 : p_j = p_j(\gamma), \quad j = 1, \ldots, m, \ \gamma \in \Gamma \subset \mathbb{R}^q.$$

(i) The test statistics

$$\chi^2_{\text{emp}} = n \sum_{j=1}^{m} \frac{(\hat{p}_j - p_j(\hat{\gamma}_n))^2}{\hat{p}_j}$$

and

$$\tilde{\chi}^2_{\text{emp}} = n \sum_{j=1}^{m} \frac{(\hat{p}_j - p_j(\hat{\gamma}_n))^2}{\hat{p}_j(\hat{\gamma}_n)}$$

are called chi-square statistics.

(ii) The tests with critical regions

$$W = \{\chi^2_{\text{emp}} \geq \chi^2_{1-\alpha}(m-1-q)\}$$

and

$$\tilde{W} = \{\tilde{\chi}^2_{\text{emp}} \geq \chi^2_{1-\alpha}(m-1-q)\}$$

are called chi-square tests.

Remark 17.11: Let N_j denote the number of observations for which the variable Y takes on the value j. The unconstrained maximum likelihood estimator $\hat{p}_j$ is given by

$$\hat{p}_j = \frac{N_j}{n}.$$

Thus the chi-square test statistics are

$$\chi^2_{\text{emp}} = \sum_{j=1}^{m} \frac{(N_j - np_j(\hat{\gamma}_n))^2}{N_j}$$

and

$$\tilde{\chi}^2_{\text{emp}} = \sum_{j=1}^{m} \frac{(N_j - np_j(\hat{\gamma}_n))^2}{np_j(\hat{\gamma}_n)}.$$

Remark 17.12: The chi-square test statistics measure how close is the empirical distribution $(\hat{p}_1, \ldots, \hat{p}_m)$ to the best approximation $(p_1(\hat{\gamma}_n), \ldots, p_m(\hat{\gamma}_n))$ of P under the null hypothesis.

17.5.2 Tests of Independence

Let (Y_i, Z_i), $i = 1, \ldots, n$, be n observations on a pair (Y, Z) of random variables. It is assumed that these pairs of observations are independent and identically distributed, that Y_i takes its values in $\{1, 2, \ldots, j, \ldots, J\}$, and that Z_i takes its values in $\{1, 2, \ldots, k, \ldots, K\}$. The common probability distribution of (Y_i, Z_i) is therefore characterized by JK positive real numbers p_{jk} satisfying $\Sigma_{j=1}^{J} \Sigma_{k=1}^{K} p_{jk} = 1$.

Let N_{jk} denote the number of observations i for which $Y_i = j$ and $Z_i = k$. The numbers N_{jk} are random and satisfy $\Sigma_{j=1}^{J} \Sigma_{k=1}^{K} N_{jk} =$

n. The joint distribution of these random numbers is the multinomial distribution $M(n; p_{jk}, j = 1, \ldots, J; k = 1, \ldots, K)$.

We want to test the null hypothesis H_0 of independence between Y_i and Z_i. Such a hypothesis says that the common distribution of (Y_i, Z_i) is the product distribution characterized by the probabilities

$$p_{jk} = p_{j\cdot} p_{\cdot k},$$

where $p_{j\cdot} = \sum_{k=1}^{K} p_{jk}$ and $p_{\cdot k} = \sum_{j=1}^{J} p_{jk}$.

The results of Section 17.4 apply with

$$\theta = (p_{jk}, (j,k) \neq (J,K))$$

and

$$\gamma = (p_{1\cdot}, \ldots, p_{J-1,\cdot}, p_{\cdot 1}, \ldots, p_{\cdot,K-1}).$$

The parameter vector θ belongs to an open subset of $\mathbb{R}^{JK-1}$ while γ belongs to an open subset of $\mathbb{R}^{J+K-2}$. The chi-square test statistics are asymptotically χ^2 distributed under H_0 with the number of degrees of freedom equal to

$$JK - 1 - J - K + 2 = (J-1)(K-1).$$

The unconstrained maximum likelihood estimator of p_{jk} is N_{jk}/n while the constrained maximum likelihood estimator is $(N_{j\cdot}N_{\cdot k})/n^2$, where

$$N_{j\cdot} = \sum_{k=1}^{K} N_{jk}, \text{ and } N_{\cdot k} = \sum_{j=1}^{J} N_{jk}.$$

By the same argument as in Section 17.5.1, it follows that we obtain the following two chi-square statistics

$$\begin{aligned} \chi^2_{\text{emp}} &= n^2 \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(\frac{N_{jk}}{n} - \frac{N_{j\cdot}N_{\cdot k}}{n^2}\right)^2}{N_{jk}} \\ &= \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(N_{jk} - \frac{N_{j\cdot}N_{\cdot k}}{n}\right)^2}{N_{jk}} \end{aligned}$$

and

$$\tilde{\chi}^2_{\text{emp}} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(N_{jk} - \frac{N_{j\cdot}N_{\cdot k}}{n}\right)^2}{N_{j\cdot}N_{\cdot k}}.$$

Definition 17.4: *Consider the problem of testing the null hypothesis of independence*

$$H_0 : p_{jk} = p_{j\cdot}p_{\cdot k},\ j = 1,\ldots,J, k = 1,\ldots,K.$$

(i) The chi-square statistics for testing such a hypothesis are given by

$$\chi^2_{\text{emp}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(N_{jk} - \frac{N_{j\cdot}N_{\cdot k}}{n}\right)^2}{N_{jk}}$$

and

$$\tilde{\chi}^2_{\text{emp}} = n\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(N_{jk} - \frac{N_{j\cdot}N_{\cdot k}}{n}\right)^2}{N_{j\cdot}N_{k\cdot}}.$$

(ii) The corresponding chi-square tests of independence are defined by the critical regions

$$W = \left\{\chi^2_{\text{emp}} \geq \chi^2_{1-\alpha}((J-1)(K-1))\right\}$$

and

$$\tilde{W} = \left\{\tilde{\chi}^2_{\text{emp}} \geq \chi^2_{1-\alpha}((J-1)(K-1))\right\}.$$

17.5.3 Goodness-of-Fit Tests to a Distribution

When the family of probability distributions associated with the null hypothesis reduces to one element $P_0 = (p_1^o, \cdots, p_m^o)$ (say), the constrained maximum likelihood estimator is clearly given by $p_i(\hat{\gamma}_n) = p_i^o$, $i = 1,\ldots,m$. Then the corresponding chi-square tests for goodness-of-fit are defined by the critical regions

$$W = \left\{\chi^2_{\text{emp}} \geq \chi^2_{1-\alpha}(m-1)\right\},\ \text{where}\quad \chi^2_{\text{emp}} = n\sum_{i=1}^{n}\frac{(\hat{p}_i - p_i^o)^2}{\hat{p}_i},$$

and

$$\tilde{W} = \left\{\tilde{\chi}^2_{\text{emp}} \geq \tilde{\chi}^2_{1-\alpha}(m-1)\right\},\ \text{where}\quad \tilde{\chi}^2_{\text{emp}} = n\sum_{i=1}^{n}\frac{(\hat{p}_i - p_i^o)^2}{p_i^o}.$$

Example 17.1: Consider the case of a discrete distribution with mass points at 0 and 1. This distribution is a Bernoulli distribution $B(1, p_1)$

for some p_1. We can use the preceding chi-square tests for testing the null hypothesis $H_0 : p_1 = p_1^o$.

Specifically, the second statistic is

$$\begin{aligned}\tilde{\chi}^2_{\text{emp}} &= n\left(\frac{(\hat{p}_0 - p_0^o)^2}{p_0^o} + \frac{(\hat{p}_1 - p_1^o)^2}{p_1^o}\right) \\ &= n\frac{(\hat{p}_1 - p_1^o)^2}{p_1^o(1-p_1^o)},\end{aligned}$$

since $p_0^o = 1 - p_1^o$ and $\hat{p}_0 = 1 - \hat{p}_1$.

17.6 Score Tests and Augmented Regressions

17.6.1 Omitted-Variable Tests in Nonlinear Regression Models

We consider the nonlinear regression model

$$y_t = m_t(\theta_1, \ldots, \theta_{p-2}) + u_t, \quad t = 1, \ldots, T, \tag{17.52}$$

where u_t are independent normal random variables with zero means and common variance θ_{p-1}. Thus the term $m_t(\theta_1, \ldots, \theta_{p-2})$ can be interpreted as the conditional expectation of y_t given some exogenous variables and possibly some lagged endogenous variables.

We want to test whether there is a specification error in the model (17.52). More precisely, we want to verify whether an additive effect due to an exogenous variable x_t has been omitted. Formally, we want to test the null hypothesis H_0 defined by the model (17.52) against the more general model

$$y_t = m_t(\theta_1, \ldots, \theta_{p-2}) + x_t\, \theta_p + u_t. \tag{17.53}$$

Thus the null hypothesis H_0 corresponds to $H_0 : \theta_p = 0$.

Within such a framework, testing H_0 can be easily carried out by means of a Wald test statistic, which is obtained by nonlinear least squares optimization. We can use the Student statistic $\hat{\theta}_p^2/\hat{\sigma}_p^2$, where $\hat{\theta}_p$ is the maximum likelihood estimator of θ_p, i.e., the nonlinear least squares estimator, and $\hat{\sigma}_p^2$ is the corresponding estimator of the variance of $\hat{\theta}_p$. The null hypothesis H_0 is rejected when this ratio is larger than the critical level $\chi^2_{1-\alpha}(1)$.

Alternatively, we can use the score test. For instance, we can use the formula $\xi_T^S = TR^2$ where R^2 is the (noncentered) coefficient of determination in the linear regression of the T-dimensional vector e whose components are all equal to one on the columns of the matrix given in Property 17.3. Such a matrix can be computed as follows.

The individual conditional log-likelihood at time t is

$$\log f_t(\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \theta_{p-1} - \frac{1}{2\theta_{p-1}}(y_t - m_t - x_t\theta_p)^2.$$

The (t,j)-th term of $\mathbf{F}$ is $\partial \log f_t(\hat{\theta}_T^0)/\partial\theta_j$, where $\hat{\theta}_T^0$ is the maximum likelihood estimator of θ under H_0. Specifically, we have

$$\begin{aligned}
\frac{\partial \log f_t(\hat{\theta}_T^0)}{\partial \theta_i} &= \frac{1}{\hat{\theta}_{p-1}^0}\frac{\partial m_t(\hat{\theta}_T^0)}{\partial \theta_i}\hat{u}_t^0, \quad i = 1,\ldots,p-2, \\
\frac{\partial \log f_t(\hat{\theta}_T^0)}{\partial \theta_{p-1}} &= -\frac{1}{2\hat{\theta}_{p-1}^0} + \frac{(\hat{u}_t^0)^2}{2(\hat{\theta}_{p-1}^0)^2}, \\
\frac{\partial \log f_t(\hat{\theta}_T^0)}{\partial \theta_p} &= \frac{1}{\hat{\theta}_{p-1}^0}x_t\hat{u}_t^0,
\end{aligned} \tag{17.54}$$

where $\hat{u}_t^0 = y_t - m_t(\hat{\theta}_T^0)$ are the residuals under H_0, $\hat{\theta}_{p-1}^0$ is the $(p-1)$-th component of $\hat{\theta}_T^0$.

In general, the score procedure is computationally more involved than the Wald procedure, which is based on nonlinear least squares. However, such a presentation of the score test procedure allows us to introduce the concept of *augmented regression*. Moreover, when applied to the problem of testing $\theta_p = 0$, equation (17.15) and the special form of $\partial \log f_t(\hat{\theta}_T^0)/\partial\theta_p$ show that the score test statistic is a function of the empirical correlation coefficient between $\hat{u}_t^0$ and the possibly omitted variable x_t. Such a correlation can be revealed graphically by considering the *residual diagram* obtained by plotting the points $(x_t, \hat{u}_t^0)$.

An example of such a diagram is given in Figure 17.2, where a positive correlation between $\hat{u}_t^0$ and x_t seems to emerge.

17.6.2 Augmented Regression Principle

Now suppose that we want to test the null hypothesis H_0 defined by (17.52) against a more general model defined by conditional densities $\tilde{f}_t(\theta_1,\ldots,\theta_{p-1},\theta_p)$ where $\tilde{f}_t(\theta_1,\ldots,\theta_{p-1},0)$ corresponds to the null hypothesis H_0. Then the null hypothesis is $H_0 : \theta_p = 0$.

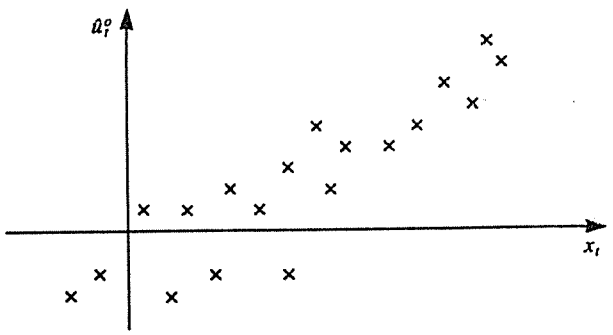


Figure 17.2: Residual Diagram

As before, the score test procedure can be applied by considering the statistic TR^2, where R^2 is now the (noncentered) coefficient of determination in the regression of the vector e on the columns of the matrix $\tilde{\mathbf{F}}$ whose generic term is $\partial \log \tilde{f}_t(\hat{\theta}_T^0)/\partial\theta_j$. It is clear that the first $p-1$ columns of $\tilde{\mathbf{F}}$ are identical to the corresponding columns of the matrix $\mathbf{F}$ defined by (17.54). Moreover, the generic term of the p-th column of $\tilde{\mathbf{F}}$ is $\partial \log \tilde{f}_t(\hat{\theta}_T^0)/\partial\theta_p$ while the corresponding generic term of $\mathbf{F}$ is $(1/\hat{\theta}_{p-1}^0)x_t\hat{u}_t^0$. Then, it is easy to see that the present score test can be interpreted as a test of omission of the variable

$$\tilde{x}_t = \frac{1}{\hat{u}_t^0} \times \frac{\partial \log \tilde{f}_t(\hat{\theta}_T^0)}{\partial\theta_p}.$$

Hence an alternative and simple way to apply the test in the general case is to consider the *augmented regression*

$$y_t = m_t(\theta_1, \ldots, \theta_{p-2}) + \theta_p\tilde{x}_t + \nu_t \tag{17.55}$$

and to compute the Student statistic associated with θ_p. It is also clear that the present argument can be generalized to the case where θ_p is multidimensional.

17.6.3 A Homoscedasticity Test

We consider now the model defined by (17.52) as a special case of the following more general model

$$y_y = m_t(\theta_1, \ldots, \theta_{p-2}) + u_t, \tag{17.56}$$

where $u_t \sim N(0, s(\theta_{p-1} + x_t\theta_p))$ and $s(\cdot)$ is a positive real function.

We want to test whether the variances of the errors are constant, i.e., whether the errors are homoscedastic. In the more general model (17.56) this corresponds to testing whether the parameter θ_p is equal to zero.

We can apply the method presented in the preceding subsection. For it is easy to see that the first $p-1$ columns of $\tilde{\mathbf{F}}$ are of the form given in (17.54) while its p-th column is

$$\frac{\partial \log \tilde{f}_t(\hat{\theta}_T^0)}{\partial \theta_p} = \frac{ds(\hat{\theta}_{p-1})}{dz} \frac{x_t}{2\hat{\sigma}_0^4}((\hat{u}_t^0)^2 - \hat{\sigma}_0^2),$$

where

$$\hat{\sigma}_0^2 = s(\hat{\theta}_{p-1}^0) = \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t^0)^2.$$

Thus we can test the null hypothesis of homoscedasticity by testing the omission of the variable

$$\frac{(\hat{u}_t^0)^2 - \hat{\sigma}_0^2}{\hat{u}_t^0} x_t.$$

Note that this test procedure does not depend on a prespecified form $s(\cdot)$ for the heteroscedasticity of the errors.

The exact form of the score test statistic can be obtained from (17.15). Specifically, we have

$$\xi_T^S = \frac{1}{2\hat{\sigma}_0^4} \frac{\left(\sum_{t=1}^{T}(x_t - \bar{x})(\hat{u}_t^0)^2\right)^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

(see Exercise 17.1). It is easy to see that the latter statistic is asymptotically equivalent to TR^2, where R^2 is the coefficient of determination in the regression of $(\hat{u}_t^0)^2$ on 1 and x_t (see Exercise 17.2).

17.6.4 A Test of Independence of the Errors

We consider the model

$$\begin{cases} y_t = m_t(\theta_1, \ldots, \theta_{p-k-1}) + u_t, \\ u_t = \theta_{p-k+1}u_{t-1} + \ldots + \theta_p u_{t-k} + \varepsilon_t, \end{cases} \tag{17.57}$$

where the errors ε_t's are independently and identically normally distributed $N(0, \theta_{p-k})$. Thus we have a nonlinear regression model whose errors follow an autoregressive process of order k.

We want to test the null hypothesis of independence of the errors u_t, i.e.

$$H_0 : \theta_{p-k+1} = \theta_{p-k+2} = \cdots = \theta_p = 0.$$

We first apply the augmented regression method. The conditional log-density function for the t-th observation is

$$\begin{aligned} \log \tilde{f}_t(\theta) &= -\frac{1}{2}\log(2\pi\theta_{p-k}) \\ &\quad -\frac{1}{2\theta_{p-k}}(y_t - m_t - \theta_{p-k+1}u_{t-1} \cdots - \theta_p u_{t-k})^2, \end{aligned}$$

where $u_{t-i} = y_{t-i} - m_{t-i}$, $i = 1, \ldots, k$. Thus we have

$$\frac{\partial \log \tilde{f}_t(\hat{\theta}_T^0)}{\partial \theta_{p-k+i}} = \frac{1}{\hat{\theta}_{p-k}^0}\hat{u}_{t-i}^0\hat{u}_t^0, \ i = 1, \ldots, k.$$

It follows that the score test can be viewed as testing the omission of the variables $\hat{u}_{t-i}^0$, $i = 1, \ldots, k$. Hence the test of H_0 can be performed by means of a Wald-type test (or a Fisher test in the linear case) in the corresponding augmented regression.

Alternatively, formulae (17.40) and (17.41) show that TR^2 can be used as a test statistic, where R^2 is the coefficient of determination in the linear regression of $\hat{u}_t^0$ on $\partial \hat{m}_t^0/\partial\theta_i$, $i = 1, \ldots, p-k-1$, and $\hat{u}_{t-j}^0$, $j = 1, \ldots, k$. The critical region of this test is $\{TR^2 \geq \chi_{1-\alpha}^2(k)\}$.

When there is no lagged endogenous variable, one can also verify that the score test statistic is asymptotically equivalent to the statistic $T(\hat{r}_1^2 + \cdots + \hat{r}_k^2)$ (see Exercise 17.4), where $\hat{r}_j$ denotes the empirical correlation coefficient between $\hat{u}_t^0$ and $\hat{u}_{t-j}^0$ when the empirical mean of $\hat{u}_t^0$ is zero. This is the case when there is a constant in the regression.

We stress however that the latter result does not hold when there are some lagged endogenous variables. In particular, when the model is linear and $k = 1$, it is easy to verify that the score test statistic is

$$\xi_T^S = \frac{T((\hat{u}^0)'(\hat{u}_{-1}^0))^2}{(\hat{u}^0)'(\hat{u}^0)((\hat{u}_{-1}^0)'(\hat{u}_{-1}^0) - (\hat{u}_{-1}^0)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{u}_{-1}^0)},$$

where $\hat{u}^0$ is the vector whose components are $\hat{u}_t^0$, $\hat{u}_{-1}^0$ is the vector whose components are $\hat{u}_{t-1}^0$, and $\mathbf{X}$ is the matrix of observations on the explanatory variables (see Exercise 17.5). When $\mathbf{X}$ does not include lagged endogenous variables, the term $(1/T)\hat{u}_{-1}^{0\prime}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{u}_{-1}^0$ converges to zero and is therefore negligible relative to $(1/T)\hat{u}_{-1}^{0\prime}\hat{u}_{-1}^0$, which converges

to the variance of the errors. In this case, as noted above, we have the asymptotic equivalence between ξ_T^S and

$$\frac{T(\hat{u}^{0\prime}\hat{u}_{-1}^0)^2}{\hat{u}^{0\prime}\hat{u}^0(\hat{u}_{-1}^{0\prime}\hat{u}_{-1}^0)} = T\hat{r}_1^2,$$

if there is a constant in the regression.

The well-known *Durbin–Watson* statistic is defined by

$$DW = \frac{\sum_{t=1}^T(\hat{u}_t^0 - \hat{u}_{t-1}^0)^2}{\sum_{t=0}^T(\hat{u}_t^0)^2}. \tag{17.58}$$

When there is no lagged endogenous variable, the score test statistic can be expressed as a simple function of the Durbin–Watson statistic, namely

$$\xi_T^S \# T\left(1 - \frac{DW}{2}\right)^2.$$

The critical region of this test is

$$\left\{T\left(1 - \frac{DW}{2}\right)^2 \geq \chi_{1-\alpha}^2(1)\right\},$$

i.e.

$$\left\{DW \leq 2\left(1 - \frac{z_{1-\alpha/2}}{\sqrt{T}}\right)\right\} \cup \left\{DW \geq 2\left(1 + \frac{z_{1-\alpha/2}}{\sqrt{T}}\right)\right\},$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-th quantile of the standard normal distribution.

17.7 Score Tests and Generalized Residuals

17.7.1 Latent Exponential Model and Score Vector

First, assume that Y_i^* is fully observed. In Chapter 11 we introduced the concept of residuals in the exponential model with conditional densities for y_i^* given x_i of the form

$$f(y_i^* \mid x_i;\theta) = \exp\left(\sum_{j=1}^r Q_j(x_i,\theta)T_j(y_i^*) + A(x_i,\theta) + B(x_i,y_i^*)\right), \tag{17.59}$$

for $i = 1, \ldots, n$ (see Section 11.3, in particular). To do so, we considered the regression model associated with the canonical statistic of the model. This regression model is

$$\begin{cases} T(Y_i^*) = m(x_i, \theta) + u_i, \\ \text{where } E(u_i \mid x_i) = 0, \\ V(u_i \mid x_i) = \omega(x_i, \theta), \end{cases} \tag{17.60}$$

and $T(Y_i^*)$ is the vector whose components are $T_j(Y_i^*)$, $j = 1, \ldots, r$. Then the residuals in this model are defined as

$$\hat{u}_i = T(Y_i^*) - m(x_i, \hat{\theta}_n),$$

where $\hat{\theta}_n$ is the maximum likelihood estimator of θ.

When a null hypothesis H_0 about θ is under test, it is useful to consider the *residuals under the null hypothesis*. They are defined as

$$\hat{u}_i^0 = T(Y_i^*) - m(x_i, \hat{\theta}_n^0), \tag{17.61}$$

where $\hat{\theta}_n^0$ is the maximum likelihood estimator of θ under H_0. Such residuals appear naturally in the score vector $\partial L_n^*(\hat{\theta}_n^0)/\partial\theta$ since

$$\frac{\partial L_n^*(\hat{\theta}_n^0)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial Q'(x_i, \hat{\theta}_n^0)}{\partial \theta} \hat{u}_i^0$$

(see Section 11.3).

Thus, if Y_i^* is observed, then we clearly have the following property.

Property 17.13: *The score test of a hypothesis about the parameter vector of an exponential model is based on the empirical covariance between $\hat{u}_i^0$ and the variables $\partial Q'(x_i, \hat{\theta}_n^0)/\partial\theta$.*

Such a property, however, can be generalized to the case where Y_i^* is incompletely observed. Specifically, assume that the observed endogenous variables Y_i are transformations of the latent variables Y_i^*, i.e.

$$Y_i = h(Y_i^*),$$

where $h(\cdot)$ is a known mapping.

We define the *generalized residuals* as

$$\tilde{u}_i = E_{\hat{\theta}_n}(u_i \mid Y_i)$$

(see Definition 11.9). Similarly, the generalized residuals under the null hypothesis are defined as H_0

$$\tilde{u}_i^0 = E_{\hat{\theta}_n^0}(u_i \mid Y_i). \tag{17.62}$$

Then the score vector under the null hypothesis H_0 is

$$\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial Q'(x_i, \hat{\theta}_n^0)}{\partial \theta} \tilde{u}_i^0. \tag{17.63}$$

This shows that here the role of the generalized residuals is similar to that of the usual residuals in the case where the latent model and the observable model are identical.

17.7.2 Application to the Probit Model

The latent linear model is

$$Y_i^* = x_i\theta + u_i, \tag{17.64}$$

where x_i is a p-dimensional row vector and the errors u_i's are independently and identically distributed $N(0,1)$. The observed endogenous variables are $Y_i = \mathbb{1}_{Y_i^*>0}$.

The density of the latent model is

$$f(y_i^* \mid x_i; \theta) = (2\pi)^{-1/2} \exp\left(x_i\theta y_i^* - \frac{1}{2}(x_i\theta)^2 - \frac{1}{2}y_i^{*2}\right).$$

Thus we have $r = 1$, $Q(x_i, \theta) = x_i\theta$ and $T(y_i^*) = y_i^*$. Hence the score vector under the null hypothesis H_0 is

$$\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} = \sum_{i=1}^{n} x_i' \tilde{u}_i,$$

where

$$\begin{aligned} \tilde{u}_i^0 &= E_{\hat{\theta}_n^0}(u_i \mid Y_i) \\ &= \frac{\phi(x_i, \hat{\theta}_n^0)}{\Phi(x_i\hat{\theta}_n^0)(1 - \Phi(x_i\hat{\theta}_n^0))} \left(y_i - \Phi(x_i\hat{\theta}_n^0)\right) \end{aligned}$$

(see Example 11.6).

The score test assesses the empirical correlation between the explanatory variables x_i and the generalized residuals under H_0. Then the score test statistic for testing whether the p-th component of x_i must be omitted is

$$\xi_n^S = \frac{(\sum_{i=1}^n x_{ip}\tilde{u}_i^0)^2}{\sum_{i=1}^n (x_{ip}\tilde{u}_i^0)^2 - \sum_{i=1}^n x_{ip}\tilde{x}_i(\tilde{u}_i^0)^2(\sum_{i=1}^n \tilde{x}_i\tilde{x}_i'(\tilde{u}_i^0)^2)^{-1}\sum_{i=1}^n x_{ip}\tilde{x}_i'(\tilde{u}_i^0)^2}, \tag{17.65}$$

where $\tilde{x}_i$ is the subvector consisting of the first $p-1$ components of x_i and x_{ip} is the p-th component of x_i (see Exercise 17.6).

17.8 Neyman or $C(\alpha)$ Tests

17.8.1 Principle

In this and the next sections we assume that the null hypothesis is of the special form H_0: $\theta_1 = \theta_1^0$, where θ_1 is a r-dimensional subvector of θ and θ_1^0 is a given r-dimensional vector. Without loss of generality, θ_1 is assumed to consist of the first r components of θ.

In such a context we have seen that the Wald statistic is

$$\xi_n^W = n(\hat{\theta}_{1n} - \theta_1^0)'(\mathcal{I}^{11}(\hat{\theta}_n))^{-1}(\hat{\theta}_{1n} - \theta_1^0), \tag{17.66}$$

while the score test statistic is

$$\xi_n^S = \frac{1}{n}\frac{\partial L_n'(\hat{\theta}_n^0)}{\partial\theta_1}\mathcal{I}^{11}(\hat{\theta}_n^0)\frac{\partial L_n(\hat{\theta}_n^0)}{\partial\theta_1}, \tag{17.67}$$

where $\hat{\theta}_{1n}$ and $\hat{\theta}_n$ are the unconstrained maximum likelihood estimators of θ_1 and θ, respectively whereas $\hat{\theta}_n^0$ is the constrained maximum likelihood estimator of θ. The likelihood ratio test statistic remains equal to

$$\xi_n^R = 2(L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0)).$$

We have also indicated that, because these three test statistics are asymptotically equivalent, a choice from them can rely upon the relative computational complexity of $\hat{\theta}_n$, $\hat{\theta}_n^0$, and $\hat{I}^{11}$. Irrespective of which test statistic is chosen, however, it is necessary to compute at least one of the two maximum likelihood estimators $\hat{\theta}_n$ and $\hat{\theta}_n^0$. It may happen that the

computation of either one of these estimators is difficult. Then it may be useful to have an alternative testing procedure that is asymptotically equivalent to the preceding three procedures and that does not require the computation of either $\hat{\theta}_n$ or $\hat{\theta}_n^0$. The $C(\alpha)$ test, which was proposed by Neyman, satisfies such a property.

Let θ_2 denote the subvector of θ such that $\theta = (\theta_1', \theta_2')'$. We assume the existence of an estimator $\bar{\theta}_{2n}$ of θ_2 such that $\sqrt{n}(\bar{\theta}_{2n} - \theta_2^0)$ converges in distribution under H_0.

Property 17.14: *Under the preceding assumptions, Neyman $C(\alpha)$ test statistic, which is defined as*

$$\begin{aligned}\xi_n^N &= \frac{1}{n}\left(\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta_1} - \mathcal{I}_{12}(\bar{\theta}_n)\mathcal{I}_{22}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta_2}\right)' \mathcal{I}^{11}(\bar{\theta}_n) \\ &\quad \times \left(\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta_1} - \mathcal{I}_{12}(\bar{\theta}_n)\mathcal{I}_{22}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta_2}\right) \\ &= \frac{1}{n}\left(\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta'}\mathcal{I}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta} - \frac{\partial L_n(\bar{\theta}_n)}{\partial \theta_2'}(\bar{\theta}_n)\mathcal{I}_{22}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta_2}\right),\end{aligned}$$

where $\bar{\theta}_n' = (\theta_1^{0\prime}, \bar{\theta}_{2n}')$, is asymptotically equivalent to the classical test statistics ξ_n^W, ξ_n^S, and ξ_n^R under H_0.

PROOF: We consider the first-order Taylor expansions of

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} \text{ and } \frac{1}{\sqrt{n}}\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta}$$

around θ_0. Such expansions give

$$0 = \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0), \tag{17.68}$$

and

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\bar{\theta}_n - \theta_0). \tag{17.69}$$

Taking the difference gives

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta} \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \bar{\theta}_n). \tag{17.70}$$

Multiplying (17.70) by the inverse of $\mathcal{I}(\theta_0)$ and partitioning this inverse give

$$\sqrt{n}(\hat{\theta}_{1n} - \theta_1^0) \# \left(\mathcal{I}^{11}(\theta_0), \mathcal{I}^{12}(\theta_0)\right)\frac{1}{\sqrt{n}}\frac{\partial L_n(\bar{\theta}_n)}{\partial \theta}.$$

Using $\mathcal{I}^{12} = -\mathcal{I}^{11}\,\mathcal{I}_{12}\,\mathcal{I}_{22}^{-1}$, we obtain

$$\sqrt{n}(\hat{\theta}_{1n} - \theta_1^0) \# \mathcal{I}^{11}(\theta_0)\frac{1}{\sqrt{n}}\left(\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_1} - \mathcal{I}_{12}(\theta_0)\mathcal{I}_{22}^{-1}(\theta_0)\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_2}\right).$$

Therefore the Wald statistic given in (17.66) is asymptotically equivalent to the Neyman statistic

$$\begin{aligned}\xi_n^N &= \frac{1}{n}\left(\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_1} - \mathcal{I}_{12}(\bar{\theta}_n)\mathcal{I}_{22}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_2}\right)' \\ &\quad \times\mathcal{I}^{11}(\bar{\theta}_n)\left(\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_1} - \mathcal{I}_{12}(\bar{\theta}_n)\mathcal{I}_{22}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_2}\right). \quad (17.71)\end{aligned}$$

Using the partitioned inverse formula (see Section A.2.2), we obtain

$$\xi_n^N = \frac{1}{n}\left(\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta'}\mathcal{I}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta} - \frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_2'}\mathcal{I}_{22}^{-1}(\bar{\theta}_n)\frac{\partial L_n(\bar{\theta}_n)}{\partial\theta_2}\right). \quad (17.72)$$

□

Remark 17.13: When the estimator $\bar{\theta}_{2n}$ coincides with the constrained maximum likelihood estimator $\hat{\theta}_{2n}^0$, then $\partial L_n(\bar{\theta}_n)/\partial\theta_2 = 0$. Thus the corresponding Neyman $C(\alpha)$ test coincides with the score test.

Remark 17.14: Suppose that $\mathcal{I}(\bar{\theta}_n)$ is replaced by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log f(y_i \mid x_i;\bar{\theta}_n)}{\partial\theta}\frac{\partial\log f(y_i \mid x_i;\bar{\theta}_n)}{\partial\theta'},$$

while $\mathcal{I}_{22}(\bar{\theta}_n)$ is replaced by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log f(y_i \mid x_i;\bar{\theta}_n)}{\partial\theta_2}\frac{\partial\log f(y_i \mid x_i;\bar{\theta}_n)}{\partial\theta_2'}.$$

Then it is easy to show that ξ_n^N can be written as

$$\xi_n^N = n(R^2 - R_2^2).$$

Here, R^2 is the (noncentered) coefficient of determination in the regression of the vector e on the columns of the $n \times p$ matrix whose generic term is $\partial\log f(y_i \mid x_i;\bar{\theta}_n)/\partial\theta_j$, while R_2^2 is the (noncentered) coefficient of determination in the regression of e on the columns of the $n \times p_2$ submatrix whose generic term is $\partial\log f(y_i \mid x_i;\bar{\theta}_n)/\partial\theta_{2j}$ with θ_{2j} being the j-th components of θ_2.

17.8.2 Example

Consider the linear model with heteroscedastic errors

$$Y_i = x_i b + u_i, \quad i = 1, \ldots, n,$$

where the conditional distribution of u_i given x_i and z_i is the normal distribution $N(0, \sigma_i^2)$ with $\sigma_i^2 = a_0 + z_i a = a_0 + z_{1i} a_1 + z_{2i} a_2$.

The null hypothesis to be tested is H_0: $a_2 = 0$. Then the parameters can be estimated by the (conditional) maximum likelihood method, unconstrained or constrained by H_0. This requires to solve relatively complex optimization problems. On the other hand, it is easy to construct consistent and asymptotically normal estimators of b, a_0, and a_1 under H_0. For b, it suffices to take the OLS estimator $\hat{b}_n$. For a_0, and a_1, it suffices to take the estimators obtained from a regression of the vector of squared residuals $\hat{u}_i^2 = (Y_i - x_i \hat{b}_n)^2$ on the subspace of dimension 2 generated by the vector e whose components are all equal to one and the vector z_1 whose components are z_{1i}, $i = 1, \ldots, n$.

Then the corresponding Neyman $C(\alpha)$ test statistic can be computed by estimating the Fisher information matrix $\mathcal{I}$ by the cross products of the first partial derivatives of

$$\log f(y_i \mid x_i; \theta) = cst - \frac{1}{2} \sum_{i=1}^{n} \log \sigma_i^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - x_i b)^2}{\sigma_i^2}$$

with respect to b, a_0, a_1, and a_2, where these parameters are evaluated at the preceding estimates for b, a_0, a_1, and at 0 for a_2.

17.9 Hausman (Likelihood Based) Tests

17.9.1 Test Description

As in the preceding section, it is assumed that the null hypothesis is of the form H_0: $\theta_1 = \theta_1^0$, where θ_1 is the vector of the first r components of θ. Let θ_2 denote the vector consisting of the $p-r$ remaining components.

The basic idea of the Hausman test is to assess the effect of imposing H_0 on the estimation of θ_2. Such an idea can be exploited within a more general framework than the present one, which is based on maximum likelihood estimation. Chapter 18 will consider such a generalization. In this chapter we consider maximum likelihood estimation only.

Specifically, we shall compare $\hat{\theta}_{2n}$ and $\hat{\theta}^0_{2n}$, which are the unconstrained and constrained estimators of θ_2 under H_0. We have the next definition and property.

Property 17.15: *The Hausman test of the null hypothesis H_0: $\theta_1 = \theta_1^0$ is the test whose critical region is*

$$\{\xi_n^H > \chi^2_{1-\alpha}(d)\},$$

where

$$\xi_n^H = n(\hat{\theta}_{2n} - \hat{\theta}^0_{2n})' \left(\mathcal{I}^{22}(\hat{\theta}^0_n) - \mathcal{I}_{22}^{-1}(\hat{\theta}^0_n)\right)^- (\hat{\theta}_{2n} - \hat{\theta}^0_{2n})$$

and d is the rank of the matrix $\mathcal{I}_{12}(\theta_0)$, which is assumed constant under H_0.

This test has asymptotic level α. Moreover, The Hausman test statistic ξ_n^H is asymptotically equivalent to ξ_n^W, ξ_n^S, ξ_n^R, and ξ^N under H_0 if and only if $d = r$.

PROOF: First, we derive the asymptotic distribution of ξ_n^H. Then we establish the stated asymptotic equivalences.

(a) ASYMPTOTIC DISTRIBUTION OF ξ_n^H:

The first-order Taylor expansions of $(1/\sqrt{n})\partial L_n(\hat{\theta}_n)/\partial\theta$ and $(1/\sqrt{n})$ $\partial L_n(\hat{\theta}^0_n)/\partial\theta$ around θ_0 give

$$0 = \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n)}{\partial\theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0)$$

and

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}^0_n)}{\partial\theta} = \begin{pmatrix} \frac{1}{\sqrt{n}}\partial L_n(\hat{\theta}^0_n)/\partial\theta_1 \\ 0 \end{pmatrix} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}^0_n - \theta_0).$$

Taking the difference gives

$$\frac{1}{\sqrt{n}}\begin{pmatrix} \partial L_n(\hat{\theta}^0_n)/\partial\theta_1 \\ 0 \end{pmatrix} \# \sqrt{n}\mathcal{I}(\theta_0)\begin{pmatrix} \hat{\theta}_{1n} - \theta_1^0 \\ \hat{\theta}_{2n} - \hat{\theta}^0_{2n} \end{pmatrix}.$$

Hence partitioning the inverse of $\mathcal{I}(\theta_0)$ gives

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{2n} - \theta^0_{2n}) \quad &\# \quad \mathcal{I}^{21}(\theta_0)\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}^0_n)}{\partial\theta_1} \\ &= \quad -\mathcal{I}_{22}^{-1}(\theta_0)\mathcal{I}_{21}(\theta_0)\mathcal{I}^{11}(\theta_0)\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}^0_n)}{\partial\theta_1} \\ &\# \quad -\mathcal{I}_{22}^{-1}(\theta_0)\mathcal{I}_{21}(\theta_0)\sqrt{n}(\hat{\theta}_{1n} - \theta_1^0). \end{aligned} \tag{17.73}$$

It follows that, under H_0, the asymptotic distribution of $\sqrt{n}(\hat{\theta}_{2n} - \hat{\theta}^0_{2n})$ is asymptotically normal, namely

$$N(0, \mathcal{I}_{22}^{-1}(\theta_0)\mathcal{I}_{21}(\theta_0)\mathcal{I}^{11}(\theta_0)\mathcal{I}_{12}(\theta_0)\mathcal{I}_{22}^{-1}(\theta_0)).$$

Using the partitioned inverse formula, we obtain an equivalent form for this asymptotic distribution, namely

$$N(0, \mathcal{I}^{22}(\theta_0) - \mathcal{I}_{22}^{-1}(\theta_0)).$$

Hence, under H_0, the Hausman test statistic

$$\xi_n^H = n(\hat{\theta}_{2n} - \hat{\theta}^0_{2n})' \left(\mathcal{I}^{22}(\hat{\theta}^0_n) - \mathcal{I}_{22}^{-1}(\hat{\theta}^0_n)\right)^- (\hat{\theta}_{2n} - \hat{\theta}^0_{2n})$$

is asymptotically distributed as a χ^2 distribution with d degrees of freedom, where d is the rank of the matrix

$$\mathcal{I}^{22}(\theta_0) - \mathcal{I}_{22}^{-1}(\theta_0) = \mathcal{I}_{22}^{-1}(\theta_0)\mathcal{I}_{21}(\theta_0)\mathcal{I}^{11}(\theta_0)\mathcal{I}_{12}(\theta_0)\mathcal{I}_{22}^{-1}(\theta_0).$$

Moreover, the last equality shows that d is equal to the rank of $\mathcal{I}_{12}(\theta_0)$.

b) ASYMPTOTIC EQUIVALENCE:

If d is equal to r, then (17.73) shows that $\sqrt{n}(\hat{\theta}_{2n} - \hat{\theta}^0_{2n})$ is obtained asymptotically from $\sqrt{n}(\hat{\theta}_{1n} - \theta^0_1)$ through an injective linear mapping. It follows that ξ_n^H and ξ_n^W are asymptotically equivalent under H_0 because these statistics are the quadratic forms associated with $\sqrt{n}(\hat{\theta}_{2n} - \hat{\theta}^0_{2n})$ and $\sqrt{n}(\hat{\theta}_{1n} - \theta^0_1)$, respectively. The remaining asymptotic equivalences follow from $\xi_n^W \# \xi_n^S \# \xi_n^R \# \xi_n^N$ under H_0.

If d is smaller than r, the asymptotic equivalence between the Hausman test and the classical test no longer holds, for the asymptotic distributions of ξ_n^H and ξ_n^W (say) have different degrees of freedom. □

First, note that the blocks of $\mathcal{I}$ or $\mathcal{I}^{-1}$ in the expression for ξ_n^H can be also evaluated at $\hat{\theta}_n$. Second, note that Property 17.15 does not depend on the choice of the generalized inverse (see Corollary B.5) provided the generalized inverse $(\mathcal{I}^{22}(\hat{\theta}^0_n) - \mathcal{I}_{22}^{-1}(\hat{\theta}^0_n))^-$ converges to $(\mathcal{I}^{22}(\theta^0) - \mathcal{I}_{22}^{-1}(\theta^0))^-$. Such a condition is important because generalized inverses are not continuous functions of the elements of the matrix that is inverted (see Property A.26). Third, note that the matrix used in the quadratic form defining ξ_n^H is the difference between the asymptotic variance covariance matrices of $\sqrt{n}(\hat{\theta}_{2n} - \theta^0_2)$ and $\sqrt{n}(\hat{\theta}^0_{2n} - \theta^0_2)$.

Lastly, as Property 17.15 states, for the Hausman test to be asymptotically equivalent to the tests introduced earlier it is necessary and

sufficient that $\mathcal{I}_{12}(\theta_2)$ is of rank r. In particular, such a condition implies the order condition $p - r \geq r$, i.e., that *the number of components of θ_2 must be equal to or greater than the number of components of θ_1.*

17.9.2 Examples

a) Application to the Linear Model

Consider the linear model

$$\begin{aligned} Y &= \mathbf{X}\theta + u \\ &= X_1\theta_1 + X_2^*\theta_{21} + X_2^{**}\theta_{22} + u, \end{aligned}$$

where X_1, X_2^*, and X_2^{**} are n-dimensional column vectors, θ_1, θ_{21}, and θ_{22} are unknown parameters, and u is a random vector distributed as $N(0, \mathbf{I})$.

The null hypothesis of interest is H_0: $\theta_1 = \theta_1^0$, which we test using the Hausman test. From (17.27), with $\mathbf{G} = (1, 0, 0)$ and $g = \theta_1^0$, we obtain

$$\begin{aligned} \hat{\theta}_{2n} - \hat{\theta}_{2n}^0 &= (\mathbf{X}'\mathbf{X})^{21}((\mathbf{X}'\mathbf{X})^{11})^{-1}(\hat{\theta}_{1n} - \theta_1^0) \\ &= (\mathbf{X}'\mathbf{X})_{22}^{-1}(\mathbf{X}'\mathbf{X})_{21}(\hat{\theta}_{1n} - \theta_1^0). \end{aligned} \tag{17.74}$$

Such an equation is the numerically exact analog of (17.73). It follows that the Hausman test statistic is

$$\xi_n^H = (\hat{\theta}_{2n} - \hat{\theta}_{2n}^0)' \left((\mathbf{X}'\mathbf{X})^{22} - (\mathbf{X}'\mathbf{X})_{22}^{-1}\right)^- (\hat{\theta}_{2n} - \hat{\theta}_{2n}^0),$$

which is numerically equal to the Wald statistic

$$\xi_n^W = (\hat{\theta}_{1n} - \theta_1^0)^2 / (\mathbf{X}'\mathbf{X})^{11}$$

provided the rank r of $(\mathbf{X}'\mathbf{X})_{21}$ is equal to one. Since $(\mathbf{X}'\mathbf{X})_{21}$ is the vector whose components are $X_2^{*\prime}X_1$ and $X_2^{**\prime}X_1$, the numerical equivalence between the Hausman test statistic and the Wald statistic holds as soon as at least one of the vectors X_2^* or X_2^{**} is not orthogonal to X_1.

If the first component of $(\mathbf{X}'\mathbf{X})^{21}$ is nonzero, then equation (17.74) shows that the Hausman test statistic reduces to

$$\xi_n^H = \frac{(\hat{\theta}_{21,n} - \hat{\theta}_{21,n}^0)^2}{V(\hat{\theta}_{21,n}) - V(\hat{\theta}_{21,n}^0)}.$$

But the first component of $(\mathbf{X}'\mathbf{X})^{21}$ is the $(2,1)$-element of $(\mathbf{X}'\mathbf{X})^{-1}$, which is

$$(\det(\mathbf{X}'\mathbf{X}))^{-1}((X_1'X_2^{**})(X_2^{*\prime}X_2^{**}) - (X_1'X_2^{*})(X_2^{**\prime}X_2^{**})).$$

In particular, even if X_2^* is orthogonal to X_1, this element can be different from zero provided X_2^{**} is not orthogonal to X_1 and X_2^*.

Similarly, if the second component of $(\mathbf{X}'\mathbf{X})^{21}$ is nonzero, then ξ_n^H becomes

$$\xi_n^H = \frac{(\hat{\theta}_{22,n} - \hat{\theta}_{22,n}^0)^2}{V(\hat{\theta}_{22,n}) - V(\hat{\theta}_{22,n}^0)}.$$

b) A Model With Autocorrelated Errors

Consider the model

$$\begin{aligned} y_t &= ay_{t-1} + bx_t + u_t, \quad |a| < 1, \\ u_t &= \rho u_{t-1} + v_t, \quad |\rho| < 1, \end{aligned}$$

for $t = 1, \ldots, T$, where the v_t's are independently and identically distributed $N(0, \sigma^2)$ and y_0 is treated as fixed. To simplify, it is assumed that σ^2 is known and equal to one. The log-likelihood function is

$$L_T = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T} v_t^2,$$

where $v_t = y_t - (\rho + a)y_{t-1} + a\rho y_{t-2} - bx_t + b\rho x_{t-1}$.

The null hypothesis of interest is the noncorrelation of the errors, i.e., H_0: $\rho = 0$. Let $\theta' = (\rho, a, b)$. Under H_0, it is easy to see that

$$\mathcal{I}(\theta_0) = \text{plim}_{T\to\infty}\left(-\frac{1}{T}\frac{\partial^2 L_T}{\partial\theta\,\partial\theta'}\right)_{\theta_0} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & \alpha & \gamma \\ 0 & \gamma & \beta \end{pmatrix},$$

where

$$\alpha = \text{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2,$$

$$\beta = \text{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} x_t^2,$$

$$\gamma = \text{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} y_{t-1} x_t.$$

Note that

$$\gamma = b \sum_{j=0}^{\infty} a^j \text{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_t x_{t-j-1}$$

is nonzero, in general. Hence $\mathcal{I}_{22}(\theta_0)$ is not diagonal. Moreover, $\mathcal{I}_{22}^{-1}(\theta_0)$ $\times \mathcal{I}_{21}(\theta_0)$, which is here equal to the first column of $\mathcal{I}_{22}^{-1}(\theta_0)$, consists of two nonzero elements. Then, from (17.73) it follows that the Hausman test can be based on either one of the components of $\hat{\theta}_2 - \hat{\theta}_2^0$, i.e., on either $\hat{a} - \hat{a}^0$ or $\hat{b} - \hat{b}^0$.

Therefore the three following test statistics are asymptotically equivalent under the null hypothesis

$$\xi^H = \begin{pmatrix} \hat{a} - \hat{a}^0 \\ \hat{b} - \hat{b}^0 \end{pmatrix}' \left(\hat{V}_{as} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} - \hat{V}_{as} \begin{pmatrix} \hat{a}^0 \\ \hat{b}^0 \end{pmatrix} \right)^{-} \begin{pmatrix} \hat{a} - \hat{a}^0 \\ \hat{b} - \hat{b}^0 \end{pmatrix}$$

$$\xi^H = (\hat{a} - \hat{a}^0)^2 / \left(\hat{V}_{as}(\hat{a}) - \hat{V}_{as}(\hat{a}^0) \right).$$

and

$$\xi^H = (\hat{b} - \hat{b}^0)^2 / \left(\hat{V}_{as}(\hat{b}) - \hat{V}_{as}(\hat{b}^0) \right),$$

From the above it follows that we can take

$$\hat{V}_{as} \begin{pmatrix} \hat{a}^0 \\ \hat{b}^0 \end{pmatrix} = \left(\sum_{t=1}^{T} (y_{t-1}, x_t)'(y_{t-1}, x_t) \right)^{-1},$$

and

$$\hat{V}_{as} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \hat{V}_{as} \begin{pmatrix} \hat{a}^0 \\ \hat{b}^0 \end{pmatrix} + \hat{V}_{as} \begin{pmatrix} \hat{a}^0 \\ \hat{b}^0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \hat{V}_{as} \begin{pmatrix} \hat{a}^0 \\ \hat{b}^0 \end{pmatrix}.$$

17.10 Exercises

EXERCISE 17.1: Consider the framework of Section 17.6.3.

a) Show that the estimators of $(\theta_1, \ldots, \theta_{p-2})$ and (θ_{p-1}, θ_p) are independent asymptotically under H_0.

b) Conclude that the Fisher information matrix is block diagonal.

c) Use (17.15) to show that the score statistic for testing homoscedasticity is

$$\frac{1}{2\hat{\sigma}_0^4}\frac{\left(\sum_{t=1}^T (x_t-\bar{x})(\hat{u}_t^0)^2\right)^2}{\sum_{t=1}^T (x_t-\bar{x})^2}.$$

EXERCISE 17.2: Consider again the homoscedasticity test of Section 17.6.3.

a) Show that the R^2 in the regression of $(\hat{u}_t^0)^2$ on 1 and x_t is equal to

$$\frac{\left(\sum_{t=1}^T (x_t-\bar{x})(\hat{u}_t^0)^2\right)^2}{\sum_{t=1}^T \left((\hat{u}_t^0)^2-\frac{1}{T}\sum_{t=1}^T (\hat{u}_t^0)^2\right)^2 \sum_{t=1}^T (x_t-\bar{x})^2}.$$

b) Using the fact that the variance of $\hat{u}_t^0$ is $2\sigma^4$ $(= 2s^2(\theta_{p-1}))$ under H_0, show that TR^2 is asymptotically equivalent to the score test statistic obtained in Exercise 17.1.

EXERCISE 17.3: COEFFICIENT STABILITY
Consider the nonlinear regression model

$$y_t = m_t(\theta_1,\ldots,\theta_{p-3}, s(\theta_{p-2}+x_t\theta_p)) + u_t,$$

where the u_t's are independently and identically distributed $N(0,\theta_{p-1})$. The null hypothesis of interest is that $s(\theta_{p-2}+x_t\theta_p)$ does not depend on t, i.e., that $\theta_p = 0$.

Show that this testing problem can be viewed as testing the omission of the variable $(\partial\hat{m}_t^0/\partial s)x_t$, where $\hat{m}_t^0 = m_t\left[\hat{\theta}_1^0,\ldots,\hat{\theta}_{p-3}^0, s(\hat{\theta}_{p-2}^0)\right]$. In particular, show that the resulting test does not depend on the form of the function s.

EXERCISE 17.4: Consider the model (17.57) and assume that there are no lagged endogenous variables. For every $j = 1,\ldots,k$, $i = 1,\ldots,p-k-1$, show that $(1/T)\sum_{t=1}^T \hat{u}_{t-j}^0 \partial\hat{m}_t^0/\partial\theta_i$ converges to zero as T goes to infinity. Using (17.40) conclude that the score test statistic is asymptotically equivalent to $T(\hat{r}_1^2+\ldots+\hat{r}_k^2)$, where $\hat{r}_j$ is the empirical correlation coefficient between $\hat{u}_t^0$ and $\hat{u}_{t-j}^0$. (Assume that the sample mean of $\hat{u}_t^0$ is zero, which is the case if there is a constant term in the regression.)

EXERCISE 17.5: Consider the model

$$\begin{cases} y_t &= \sum_{i=1}^{p-2} x_{it}\theta_i + u_i, \\ u_t &= \theta_p u_{t-1} + \epsilon_t, \end{cases}$$

where the ϵ_t's are independently and identically distributed $N(0, \theta_{p-1})$. The x_{it}'s may include lagged endogenous variables. The null hypothesis of interest is H_0: $\theta_p = 0$.

a) Determine the likelihood function (conditional upon the initial values).

b) Show that the element corresponding to (θ_{p-1}, θ_i) in the Fisher information matrix is null for each $i \neq p-1$.

c) Show that the Fisher information matrix that is obtained by omitting θ_{p-1} is

$$\frac{1}{T\hat{\theta}^0_{p-1}} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\hat{u}^0_{-1} \\ (\hat{u}^0_{-1})'\mathbf{X} & (\hat{u}^0_{-1})'(\hat{u}^0_{-1}) \end{bmatrix}$$

under H_0.

d) Using (17.15), show that

$$\xi^S_T = \frac{T\left((\hat{u}^0)'(\hat{u}^0_{-1})\right)'}{(\hat{u}^0)'(\hat{u}^0)\left((\hat{u}^0_{-1})'(\hat{u}^0_{-1}) - (\hat{u}^0_{-1})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{u}^0_{-1}\right)}$$

EXERCISE 17.6: Consider the Probit model (17.64). The null hypothesis of interest is H_0: $\theta_p = 0$. Show that the generic term of the $n \times p$ matrix $\mathbf{F}$, which is defined in Property 17.3, is $x_{ij}\tilde{u}^0_i$, for $i = 1, \ldots, n$, $j = 1, \ldots, p$. Then, using the general formula for the score test statistic given in Property 17.3, show that ξ^S_n is given by formula (17.65).

EXERCISE 17.7: Let $\hat{\theta}^0_n$ be the maximum likelihood estimator of θ under a null hypothesis, which can be either in implicit or explicit form. Show that $(\hat{\theta}^0_n - \theta_0)'\partial L_n(\hat{\theta}^0_n)/\partial\theta$ converges to zero in probability when n increases to infinity.

17.11 References

Breusch, T.S. and Pagan, A.R. (1980). "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, 239–254.

Durbin, J. and Watson, G. (1950). "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, 37, 409–428.

Durbin, J. and Watson, G. (1951). "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, 38, 159–178.

Engle, R.F. (1984). "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics," in *Handbook of Econometrics*, Vol. 2, North-Holland.

Gourieroux, C., Monfort, A., Renault, E., and Trognon, A. (1987). "Generalised Residuals ," *Journal of Econometrics*, 34, 5–32.

Hausman, J. (1978). "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1272.

Holly, A. (1982). "A Remark on Hausman's Specification Test," *Econometrica*, 50, 749–759.

Neyman, J. (1957). "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics*, Almquist and Wiksell.

Pearson, K. (1900). "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to Have Arisen from Random Sampling," *Phil. Mag.*, Ser. 5, 50, 157–172.

Silvey, S.D. (1959). "The Lagrange Multiplier Test," *The Annals of Mathematical Statistics*, 30, 389–407.

Silvey, S.D. (1970). *Statistical Inference*, Chapman and Hall.

Wald, A. (1934). "Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54, 426–442.

Wilks, S.S. (1938). "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *Annals of Mathematical Statistics*, 9, 60–62.

Wilks, S.S. (1962). *Mathematical Statistics*, Wiley.

CHAPTER 18

General Asymptotic Tests

In the previous chapter we have considered tests based on the likelihood function. Such tests can be used when the model is parametric, i.e., when the family of probability distributions defining the statistical model is indexed by a finite dimensional parameter. Moreover, we have assumed that this family contains the true probability distribution having generated the observations. When such assumptions are deemed too stringent, it is necessary to apply other testing procedures.

In Section 1 we shall consider the case where the unconstrained and constrained estimators of the parameter vector θ of interest are obtained via a maximization of a general statistical criterion $L_n(\theta)$. In addition, we shall consider the case where the null hypothesis is under the so-called *mixed form* (See Section 10.1). The resulting framework is quite general. It contains as a special case the theory developed in the previous chapter, which relied upon the likelihood function and a null hypothesis in either implicit or explicit form. In Section 2, we shall study *specification tests* based on asymptotic least squares. In Section 3 we shall consider specification tests based on generalized methods of moments. These estimation methods were introduced in Chapter 9. In Section 4 we shall apply the general principle underlying Hausman tests to the present framework. Such a principle was introduced in Section 17.9 within a likelihood context. Lastly, in Section 5 we shall present the so-called *information matrix test.*

18.1 Tests Based on an Estimation Procedure Defined by a General Objective Function

We consider the framework of Section 10.3.1. The parameter of interest is $\theta \in \Theta \subset R^p$. The unconstrained estimator $\hat{\theta}_n$ of θ is obtained as a

solution to the maximization problem

$$\max_{\theta \in \Theta} L_n(\theta). \tag{18.1}$$

It is assumed that the unconstrained estimator $\hat{\theta}_n$ satisfies the first-order conditions

$$\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} = 0, \tag{18.2}$$

that the score vector is asymptotically normally distributed, i.e.

$$\frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \mathcal{I}_0), \tag{18.3}$$

and that

$$-\frac{1}{n} \frac{\partial^2 L_n(\theta_0)}{\partial \theta \partial \theta'} \to \mathcal{J}_0, \text{ with probability 1.} \tag{18.4}$$

Here θ_0 denotes the true value of θ, and $\mathcal{I}_0$ and $\mathcal{J}_0$ are two nonstochastic symmetric positive definite matrices.

The hypothesis of interest is written under the general *mixed form*

$$H_0 : \{\theta : \exists a \in A \subset \mathbb{R}^q, g(\theta, a) = 0\}, \tag{18.5}$$

where a is an auxiliary parameter and g is a vector function with values in $\mathbb{R}^r$. As in Chapter 10, some rank conditions are imposed on the function g defining the constraints under test. These constraints are

$$\operatorname{rank}\left(\frac{\partial g}{\partial \theta'}\right) = r \quad \text{and} \quad \operatorname{rank}\left(\frac{\partial g}{\partial a'}\right) = q, \tag{18.6}$$

which implies that $q \leq r \leq p$.

We shall distinguish three kinds of tests: (i) Tests using the constrained maximization of $L_n(\theta)$, (ii) tests based only on the unconstrained estimator $\hat{\theta}_n$, and (iii) tests based on the function $g(\hat{\theta}_n, a)$.

18.1.1 Tests Using the Constrained Maximization of $L_n(\theta)$

The tests discussed here all involve (at least) the constrained maximization of the objective function $L_n(\theta)$. The constrained estimators $\hat{\theta}_n^0$

and $\hat{a}_n^0$ of θ and a respectively solve the maximization problem

$$\begin{cases} \max_{\theta,a} L_n(\theta) \\ \\ \text{subject to} \quad g(\theta, a) = 0 \end{cases} \tag{18.7}$$

We let $\hat{\lambda}_n$ denote the vector of r Lagrange multipliers associated with the constraint $g(\theta, a) = 0$.

In Chapter 10 we derived some Taylor series expansions for the estimators $\hat{\theta}_n^0$, $\hat{a}_n^0$, and $\hat{\lambda}_n^0$ under the null hypothesis. Such expansions are summarized by the system

$$\begin{pmatrix} \mathcal{J}_0 & 0 & \partial g'/\partial\theta \\ 0 & 0 & \partial g'/\partial a \\ \partial g/\partial\theta' & \partial g/\partial a' & 0 \end{pmatrix} \times \begin{pmatrix} \sqrt{n}(\hat{\theta}_n^0 - \theta_0) \\ \sqrt{n}(\hat{a}_n^0 - a_0) \\ \frac{1}{\sqrt{n}}\hat{\lambda}_n \end{pmatrix} \# \begin{pmatrix} \mathcal{J}_0\sqrt{n}(\hat{\theta}_n - \theta_0) \\ 0 \\ 0 \end{pmatrix}, \tag{18.8}$$

where all functions are evaluated at the true values θ_0 and a_0 of θ and a (see (10.33)).

In particular, this system can be solved for the quantities

$$\sqrt{n}(\hat{a}_n^0 - a_0), \quad \frac{1}{\sqrt{n}}\hat{\lambda}_n \quad \text{and} \quad \sqrt{n}(\hat{\theta}_n^0 - \theta_0)$$

as functions of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. Specifically, we obtain

$$\begin{aligned} \sqrt{n}(\hat{a}_n^0 - a_0) \quad \# \quad & -\left\{ \frac{\partial g'}{\partial a}\left(\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\right)^{-1}\frac{\partial g}{\partial a'}\right\}^{-1} \\ & \times \frac{\partial g'}{\partial a}\left(\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\right)^{-1}\frac{\partial g}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0), \end{aligned} \tag{18.9}$$

$$\frac{1}{\sqrt{n}}\hat{\lambda}_n \# \left(\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\right)^{-1}(\mathbf{I} - \mathbf{M})\frac{\partial g}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0), \tag{18.10}$$

and

$$\begin{aligned} &\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \# \\ &\left\{\mathbf{I} - \mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\left(\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\right)^{-1}(\mathbf{I} - \mathbf{M})\frac{\partial g}{\partial \theta'}\right\} \times \sqrt{n}(\hat{\theta}_n - \theta_0), \end{aligned} \tag{18.11}$$

where $\mathbf{M}$ is the projection matrix defined by

$$\mathbf{M} = \frac{\partial g}{\partial a'}\left\{\frac{\partial g'}{\partial a}\left(\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\right)^{-1}\frac{\partial g'}{\partial a}\right\}^{-1}\frac{\partial g'}{\partial a}\left(\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\right)^{-1}. \tag{18.12}$$

The matrix $\mathbf{M}$ can also be written as

$$\mathbf{M} = \frac{\partial g}{\partial a'}\left(\frac{\partial g'}{\partial a}\mathbf{\Omega}_0^{-1}\frac{\partial g}{\partial a'}\right)^{-1}\frac{\partial g'}{\partial a}\mathbf{\Omega}_0^{-1}, \tag{18.13}$$

where

$$\mathbf{\Omega}_0 = \frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}.$$

a) Wald Type Tests

By definition of the constrained estimators we have $g(\hat{\theta}_n^0, \hat{a}_n^0) = 0$. Thus it is natural to constuct a test based on the difference between 0 and $g(\hat{\theta}_n, \hat{a}_n^0)$, where $\hat{\theta}_n$ is the unconstrained estimator of θ.

Property 18.1: *Under H_0, the vector $\sqrt{n}g(\hat{\theta}_n, \hat{a}_n^0)$ converges in distribution to a multivariate centered normal distribution with variance covariance matrix given by*

$$V_{as}\left(\sqrt{n}\; g(\hat{\theta}_n, \hat{a}_n^0)\right) = (\mathbf{I} - \mathbf{M})\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\mathcal{I}_0\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}(\mathbf{I} - \mathbf{M})'.$$

Let the Wald statistic be defined as

$$\xi_n^W = ng'(\hat{\theta}_n, \hat{a}_n^0)\hat{V}_{as}^{-}\left(\sqrt{n}g(\hat{\theta}_n, \hat{a}_n^0)\right)g(\hat{\theta}_n, \hat{a}_n^0),$$

where $\hat{V}_{as}^{-}\left(\sqrt{n}g(\hat{\theta}_n, \hat{a}_n^0)\right)$ is a consistent estimator of a generalized inverse of $\hat{V}_{as}\left(\sqrt{n}g(\hat{\theta}_n, \hat{a}_n^0)\right)$. Then, ξ_n^W is asymptotically distributed as a χ^2 distribution with $r - q$ degrees of freedom under H_0. Hence, the critical region $\{\xi_n^W \geq \chi_{1-\alpha}^2(r-q)\}$ defines a test with asymptotic level equal to α.

PROOF: We have

$$\sqrt{n}g(\hat{\theta}_n, \hat{a}_n^0) \;\#\; \frac{\partial g}{\partial \theta'}\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) + \frac{\partial g}{\partial a'}\sqrt{n}(\hat{a}_n^0 - a_0).$$

Then, using (18.9) and (18.12), we obtain

$$\sqrt{n}\, g(\hat{\theta}_n, \hat{a}_n^0) \# (\mathbf{I} - \mathbf{M}) \frac{\partial g}{\partial \theta'} \sqrt{n} \left(\hat{\theta}_n - \theta_0 \right). \tag{18.14}$$

The asymptotic distribution of $\sqrt{n}\, g(\hat{\theta}_n, \hat{a}_n^0)$ follows from the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. The latter distribution is a $N(0, \mathcal{J}_0^{-1}\mathcal{I}_0\mathcal{J}_0^{-1})$ (see (10.21)). The remaining statements of the property follow immediately. □

b) Lagrange Multiplier Tests

Property 18.2: *Let ξ_n^L be the test statistic defined as*

$$\xi_n^L = (1/n)\hat{\lambda}_n' \hat{V}_{as}^- \left(\hat{\lambda}_n / \sqrt{n} \right) \hat{\lambda}_n,$$

where $\hat{V}_{as}^- \left(\hat{\lambda}_n / \sqrt{n} \right)$ is a consistent estimator of a generalized inverse of

$$V_{as}(\hat{\lambda}_n / \sqrt{n}) = \mathbf{\Omega}_0^{-1}(\mathbf{I} - \mathbf{M}) \frac{\partial g}{\partial \theta'} \mathcal{J}_0^{-1} \mathcal{I}_0 \mathcal{J}_0^{-1} \frac{\partial g'}{\partial \theta} (\mathbf{I} - \mathbf{M}') \mathbf{\Omega}_0^{-1}.$$

Then the statistics ξ_n^L and ξ_n^W are asymptotically equivalent under H_0.

PROOF: Equations (18.10) and (18.14) imply

$$\begin{aligned} \frac{1}{\sqrt{n}} \hat{\lambda}_n \quad & \# \quad \mathbf{\Omega}_0^{-1}(\mathbf{I} - \mathbf{M}) \frac{\partial g}{\partial \theta'} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ & \# \quad \mathbf{\Omega}_0^{-1} \sqrt{n} g(\hat{\theta}_n, \hat{a}_n^0). \end{aligned}$$

Using Property 18.1, this gives the asymptotic variance covariance matrix of $(1/\sqrt{n})\hat{\lambda}_n$.

The asymptotic equivalence between ξ_n^L and ξ_n^W follows from the fact that $(1/\sqrt{n})\hat{\lambda}_n$ and $\sqrt{n}\, g(\hat{\theta}_n, \hat{a}_n^0)$ are asymptotically in a linear bijective relationship. □

From Property 18.2 it follows that ξ_n^L is asymptotically distributed as a $\chi^2(r-q)$ under H_0. Thus the critical region $\{\xi_n^L \geq \chi_{1-\alpha}^2(r-q)\}$ defines a test with asymptotic level equal to α.

c) Hausman–Wald Tests

Within the general framework of this section, we can construct a test based on the difference between $\hat{\theta}_n$ and $\hat{\theta}_n^0$. Such a testing approach is

similar to both the Wald type procedure proposed in Section 17.4.1 and to the Hausman procedure presented in Section 17.9.

Property 18.3: *Let ξ_n^H be the test statistic defined as*

$$\xi_n^H = n\left(\hat{\theta}_n - \hat{\theta}_n^0\right)' \hat{V}_{as}^{-}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)\left(\hat{\theta}_n - \hat{\theta}_n^0\right),$$

where $\hat{V}_{as}^{-}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)$ is a consistent estimator of a generalized inverse of

$$V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right) = \mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}V_{as}(\hat{\lambda}_n/\sqrt{n})\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}$$

and $V_{as}(\hat{\lambda}_n/\sqrt{n})$ is given in Property 18.2. Then, under H_0, we have

$$\xi_n^H \# \xi_n^L \# \xi_n^W.$$

PROOF: From (18.11) and (18.13) it follows that

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \# \mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\Omega_0^{-1}(\mathbf{I} - \mathbf{M})\frac{\partial g}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0).$$

Hence, using (18.10), we obtain

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \# \mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}} \tag{18.15}$$

This gives the asymptotic variance covariance matrix of $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$.

Moreover, the mapping $\mathcal{J}_0^{-1}\partial g'/\partial\theta$ is injective because of the rank conditions (18.6). It follows that $\xi_n^H \# \xi_n^L$. Hence $\xi_n^H \# \xi_n^W$ by Property 18.2. □

Property 18.3 implies that ξ_n^H is asymptotically distributed as a $\chi^2(r-q)$ under H_0. Thus the critical region $\{\xi_n^H \geq \chi_{1-\alpha}^2(r-q)\}$ defines a test with asymptotic level α.

d) Score Tests

The general principle of constructing a test based on the difference between 0 and the score vector $\partial L_n(\hat{\theta}_n^0)/\partial\theta$ can also be invoked here.

Property 18.4: *Let ξ_n^S be the test statistic defined as*

$$\xi_n^S = \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\hat{V}_{as}^{-}\left(\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}\right)\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta},$$

where $\hat{V}_{as}^{-}\left((1/\sqrt{n})(\partial L_n(\hat{\theta}_n^0)/\partial\theta)\right)$ is a consistent estimator of a generalized inverse of

$$V_{as}\left(\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}\right) = \mathcal{J}_0 V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)\mathcal{J}_0$$

and $V_{as}(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0))$ is given in Property 18.3. Then, under H_0, we have

$$\xi_n^S \# \xi_n^H \# \xi_n^L \# \xi_n^W.$$

PROOF: Expanding the score vector $(1/\sqrt{n})(\partial L_n(\hat{\theta}_n^0)/\partial\theta)$ in a Taylor series around θ_0 yields

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} + \frac{1}{n}\frac{\partial^2 L_n(\theta_0)}{\partial\theta\partial\theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0),$$

i.e.

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{J}_0\sqrt{n}(\hat{\theta}_n^0 - \theta_0). \tag{18.16}$$

On the other hand, the unconstrained estimator satisfies the first-order conditions

$$0 = \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta^0)}{\partial \theta} - \mathcal{J}_0\sqrt{n}\left(\hat{\theta}_n - \theta_0\right),$$

or equivalently

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} \# \mathcal{J}_0\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{18.17}$$

Substituting this expression for the score vector in (18.16) yields

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} \# \mathcal{J}_0\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0).$$

The asymptotic variance covariance matrix of $(1/\sqrt{n})(\partial L_n(\hat{\theta}_n^0)/\partial\theta)$ follows. Moreover, the asymptotic equivalence between ξ_n^S and ξ_n^H is obtained from the nonsingularity of $\mathcal{J}_0$. □

From Property 18.4 it follows that ξ_n^S is asymptotically distributed as a $\chi^2(r-q)$ under H_0. Thus the critical region $\{\xi_n^S \geq \chi^2_{1-\alpha}(r-q)\}$ defines a test with asymptotic level equal to α.

e) Tests Based on the Difference Between the Maximal Values of the Objective Function

In Chapter 17 we showed that the test statistic $2(L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0))$ is asymptotically equivalent to the preceding test statistics when the objective function L_n is the log-likelihood function. It is natural to ask whether this result still holds in the general framework considered here. Unlike the four preceding tests, such an asymptotic equivalence does not always hold. It holds, however, in the important special case where $\mathcal{I}_0 = \mathcal{J}_0$. Recall that the latter equality is satisfied not only for maximum likelihood estimation but also for other estimation methods such as quasi generalized pseudo maximum likelihood, optimal asymptotic least squares and optimal generalized methods of moments (see Chapter 10).

Property 18.5: *If $\mathcal{I}_0 = \mathcal{J}_0$, then the test statistic*

$$\xi_n^M = 2\left(L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0)\right)$$

is asymptotically equivalent to ξ_n^W, ξ_n^L, ξ_n^H, and ξ_n^S under H_0.

PROOF: As in the proof of Property 17.4, we take a Taylor series expansion of ξ_n^M under the null hypothesis. This gives

$$\xi_n^M \# n(\hat{\theta}_n - \hat{\theta}_n^0)'\mathcal{J}_0(\hat{\theta}_n - \hat{\theta}_n^0). \tag{18.18}$$

In the general case where $\mathcal{I}_0 \neq \mathcal{J}_0$ this statistic is not asymptotically χ^2 distributed.

When $\mathcal{I}_0 = \mathcal{J}_0$, however, we shall show that $\mathcal{J}_0$ is a generalized inverse of $V_{as}(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0))$. Using Property 18.3, this implies the asymptotic equivalence between ξ_n^M and ξ_n^H and hence that $\xi_n^M \# \xi_n^H \# \xi_n^W \# \xi_n^L \# \xi_n^S$.

From Properties 18.2 and 18.3 we have

$$V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right) =$$
$$\mathcal{J}_0^{-1}\frac{\partial g'}{\partial\theta}\Omega_0^{-1}(\mathbf{I}-\mathbf{M})\frac{\partial g}{\partial\theta'}\mathcal{J}_0^{-1}\mathcal{I}_0\mathcal{J}_0^{-1}\frac{\partial g'}{\partial\theta}(\mathbf{I}-\mathbf{M}')\Omega_0^{-1}\frac{\partial g}{\partial\theta'}\mathcal{J}_0^{-1}.$$

When $\mathcal{I}_0 = \mathcal{J}_0$, this expression simplifies and we obtain

$$V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right) = \mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\boldsymbol{\Omega}_0^{-1}(\mathbf{I}-\mathbf{M})\boldsymbol{\Omega}_0(\mathbf{I}-\mathbf{M}')\boldsymbol{\Omega}_0^{-1}\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}.$$

It remains to show that $\mathcal{J}_0$ is a generalized inverse of $V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)$. This follows from the next lemma.

Lemma 18.1: *If* $\mathcal{I}_0 = \mathcal{J}_0$, *then*

$$V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)\mathcal{J}_0 V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right) = V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right).$$

PROOF: We have

$$V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)\mathcal{J}_0 V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right) =$$

$$\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\boldsymbol{\Omega}_0^{-1}(\mathbf{I}-\mathbf{M})\boldsymbol{\Omega}_0(\mathbf{I}-\mathbf{M}')\boldsymbol{\Omega}_0^{-1}(\mathbf{I}-\mathbf{M})\boldsymbol{\Omega}_0(\mathbf{I}-\mathbf{M}')\boldsymbol{\Omega}_0^{-1}\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}. \tag{18.19}$$

Since $\mathbf{I}-\mathbf{M}$ is an orthogonal projection matrix with respect to the scalar product defined by $\boldsymbol{\Omega}_0^{-1}$, we have $(\mathbf{I}-\mathbf{M}')\boldsymbol{\Omega}_0^{-1}\mathbf{M} = 0$, or equivalently $(\mathbf{I}-\mathbf{M})\boldsymbol{\Omega}_0\mathbf{M}' = 0$. The latter equality implies

$$\begin{aligned}
&V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)\mathcal{J}_0 V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right) \\
&\quad = \mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\boldsymbol{\Omega}_0^{-1}(\mathbf{I}-\mathbf{M})\boldsymbol{\Omega}_0\boldsymbol{\Omega}_0^{-1}(\mathbf{I}-\mathbf{M})\boldsymbol{\Omega}_0(\mathbf{I}-\mathbf{M}')\boldsymbol{\Omega}_0^{-1}\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1} \\
&\quad = \mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\boldsymbol{\Omega}_0^{-1}(\mathbf{I}-\mathbf{M})\boldsymbol{\Omega}_0(\mathbf{I}-\mathbf{M}')\boldsymbol{\Omega}_0^{-1}\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1} \\
&\quad = V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right).
\end{aligned}$$

□

f) Simplification of the Test Statistics When $\mathcal{I}_0 = \mathcal{J}_0$

When $\mathcal{I}_0 = \mathcal{J}_0$, some simplifications arise in the expressions of the test statistics ξ_n^W, ξ_n^L, ξ_n^H, and ξ_n^S introduced earlier.

Property 18.6: *If* $\mathcal{I}_0 = \mathcal{J}_0$, *then*

$$\begin{aligned}
\xi_n^W &= n g'(\hat{\theta}_n, \hat{a}_n^0)\hat{\Omega}_n^{-1} g(\hat{\theta}_n, \hat{a}_n^0), \\
\xi_n^L &= \frac{1}{n}\hat{\lambda}_n'\hat{\Omega}_n\hat{\lambda}_n, \\
\xi_n^H &= n(\hat{\theta}_n - \hat{\theta}_n^0)\hat{\mathcal{J}}_n(\hat{\theta}_n - \hat{\theta}_n^0), \\
\xi_n^S &= \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\hat{\mathcal{J}}_n^{-1}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta},
\end{aligned}$$

where $\hat{\Omega}_n$ *and* $\hat{\mathcal{J}}_n$ *are consistent estimators of* Ω_0 *and* $\mathcal{J}_0$ *respectively.*

PROOF:

(i) If $\mathcal{I}_0 = \mathcal{J}_0$, then Property 18.1 implies

$$V_{as}\left(\sqrt{n} g(\hat{\theta}_n, \hat{a}_n^0)\right) = (\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}').$$

But Ω_0^{-1} is a generalized inverse of $(\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}')$ because

$$\begin{aligned}
&(\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}')\Omega_0^{-1}(\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}') \\
&\quad = (\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}')\Omega_0^{-1}\Omega_0(\mathbf{I} - \mathbf{M}') \\
&\quad = (\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}'),
\end{aligned}$$

where we have used the equality $(\mathbf{I} - \mathbf{M}')\Omega_0^{-1}\mathbf{M} = 0$. The expression for ξ_n^W follows.

(ii) If $\mathcal{I}_0 = \mathcal{J}_0$, then Property 18.2 implies

$$V_{as}(\hat{\lambda}_n/\sqrt{n}) = \Omega_0^{-1}(\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}')\Omega_0^{-1}.$$

But Ω_0 is a generalized inverse of such a variance covariance matrix because Ω_0^{-1} is a generalized inverse of $(\mathbf{I} - \mathbf{M})\Omega_0(\mathbf{I} - \mathbf{M}')$. The expression for ξ_n^L follows.

(iii) When $\mathcal{I}_0 = \mathcal{J}_0$, $\mathcal{J}_0$ is a generalized inverse of $V_{as}(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0))$ (see Lemma 18.1). The expression for ξ_n^H follows.

(iv) Lastly, from the proof of Property 18.4 we obtain

$$V_{as}\left(\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}\right) = \mathcal{J}_0 V_{as}\left(\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)\right)\mathcal{J}_0.$$

If $\mathcal{I}_0 = \mathcal{J}_0$, Lemma 18.1 implies that $\mathcal{J}_0^{-1}$ is a generalized inverse of $V_{as}((1/\sqrt{n})(\partial L_n(\hat{\theta}_n^0)/\partial\theta))$. The expression for ξ_n^S follows. □

The expressions for the test statistics ξ_n^W, ξ_n^L, ξ_n^H, and ξ_n^S obtained in Property 18.6 are quite simple. In particular, the latter two expressions present the convenient feature that the matrix $\mathcal{J}_n$ (or $\mathcal{J}_n^{-1}$) defining the two corresponding quadratic forms does not depend on the null hypothesis under test, provided the maintained hypothesis does not change. In addition, the latter two statistics involve $\mathcal{J}_n$, which is nonsingular because $\mathcal{I}_0$ is nonsingular. Hence there is no need to use generalized inverses, which may raise some discontinuity problems when matrices are singular (see Property A.26).

g) Exact or Asymptotic Invariance of the Test Statistics With Respect to the Constraints Defining H_0

Suppose that the null hypothesis H_0 is defined by another function $h(\theta, b)$ satisfying

$$\{\theta : \exists a, g(\theta, a) = 0\} = \{\theta : \exists b, h(\theta, b) = 0\}.$$

The test statistics ξ_n^H and ξ_n^S remain unchanged since they do not depend on how the null hypothesis is defined. On the other hand, the test statistics ξ_n^W and ξ_n^L will typically change. The new test statistics associated with the defining function h are, however, asymptotically equivalent to the test statistics ξ_n^H and ξ_n^S and hence to the original test statistics ξ_n^W and ξ_n^L. In addition, when $\mathcal{I}_0 = \mathcal{J}_0$, it is easy to verify that $\xi_n^L = \xi_n^S$. Thus ξ_n^L is invariant with respect to the constraints defining H_0.

18.1.2 Tests Based on the Unconstrained Estimator $\hat{\theta}_n$ of θ

The preceding five tests all involve the constrained maximization of the objective function $L_n(\theta)$. It is possible to replace such a maximization by the constrained maximization of a quadratic form depending on the observations through $\hat{\theta}_n$ only.

Specifically, consider the constrained estimator $(\tilde{\theta}_n^0, \tilde{a}_n^0)$ of (θ, a) defined as a solution to

$$\begin{cases} \min_{\theta,a} (\hat{\theta}_n - \theta)' \mathbf{W}_n (\hat{\theta}_n - \theta) \\ \text{subject to} \quad g(\theta, a) = 0, \end{cases} \tag{18.20}$$

or equivalently, as a solution to

$$\begin{cases} \max_{\theta,a} \Psi_n(\theta) = -\frac{n}{2}(\hat{\theta}_n - \theta)'\mathbf{W}_n(\hat{\theta}_n - \theta) \\ \text{subject to} \quad g(\theta, a) = 0, \end{cases} \tag{18.21}$$

where $\{\mathbf{W}_n\}$ is a sequence of matrices converging to a symmetric positive definite matrix $\mathbf{W}_0$. Let $\tilde{\lambda}_n$ denote the vector of Lagrange multipliers associated with the above constraints.

We have the following property.

Property 18.7: *If $\mathbf{W}_0 = \mathcal{J}_0$, then the Wald, Lagrange multiplier, Hausman, and score test statistics associated with the objective function $\Psi_n(\theta)$ defined in (18.21) are asymptotically equivalent , under H_0, to the test statistics based on the objective function $L_n(\theta)$ defined in Section 18.1.1.*

PROOF: The unconstrained maximum of $\Psi_n(\theta)$ is clearly attained when $\theta = \hat{\theta}_n$. In addition, from (18.21) we have

$$\frac{1}{\sqrt{n}} \frac{\partial \Psi_n(\theta_0)}{\partial \theta} = \mathbf{W}_n \sqrt{n}(\hat{\theta}_n - \theta_0) \# \mathbf{W}_0 \sqrt{n}(\hat{\theta}_n - \theta_0). \tag{18.22}$$

Hence $(1/\sqrt{n})(\partial \Psi_n(\theta_0)/\partial \theta)$ is asymptotically normally distributed. In addition, $-(1/n)(\partial^2 \Psi_n(\theta_0)/\partial \theta \partial \theta') = \mathbf{W}_n$, which converges to $\mathbf{W}_0$. As a result, the objective function Ψ_n satisfies the analogs of the conditions (18.3) and (18.4). Thus $\tilde{\theta}_n^0$, $\tilde{a}_n^0$, and $\tilde{\lambda}_n$ satisfy a system similar to (18.8), where $\mathcal{J}_0$ is replaced by $\mathbf{W}_0$. Hence, when $\mathbf{W}_0 = \mathcal{J}_0$, these two systems are identical. The desired asymptotic equivalences follow. □

In the important case where $\mathcal{I}_0 = \mathcal{J}_0$, we obtain an even more interesting result.

Property 18.8: *If $\mathbf{W}_0 = \mathcal{J}_0 = \mathcal{I}_0$, then all the test statistics of Property 18.7 are asymptotically equivalent under H_0 to*

$$-2\Psi_n(\tilde{\theta}_n^0) = n(\hat{\theta}_n - \tilde{\theta}_n^0)'\mathbf{W}_n(\hat{\theta}_n - \tilde{\theta}_n^0).$$

PROOF: If $\mathcal{J}_0 = \mathbf{W}_0$, then the asymptotic variance covariance matrix of $(1/\sqrt{n})(\partial \Psi_n(\theta_0)/\partial \theta)$ is $\mathcal{J}_0 \mathcal{J}_0^{-1} \mathcal{I}_0 \mathcal{J}_0^{-1} \mathcal{J}_0 = \mathcal{I}_0$ using (18.22) and (10.21). Moreover, since $\mathcal{I}_0 = \mathcal{J}_0$, then we can apply Property 18.5 to the objective function Ψ_n. Thus we obtain the asymptotically equivalent test statistic

$$2\left(\Psi_n(\hat{\theta}_n) - \Psi_n(\tilde{\theta}_n^0)\right) = -2\Psi_n(\tilde{\theta}_n^0).$$

□

Property 18.8 can be used to obtain directly a test statistic for testing H_0 that is asymptotically equivalent to the preceding test statistics and that relies upon the optimal value of the objective function Ψ_n under the constraints $g(\theta, a) = 0$.

18.1.3 Tests Based on the Function $g(\hat{\theta}_n, a)$

Other asymptotically equivalent tests can be constructed by assessing whether the constraint defining the null hypothesis is approximately satisfied by the unconstrained estimator $\hat{\theta}_n$. This idea leads to the introduction of an appropriate measure of the discrepancy between $g(\hat{\theta}_n, a)$ and zero.

Consider the minimization problem

$$\min_a \Phi_n(a) = ng'(\hat{\theta}_n, a)\mathbf{S}_n g(\hat{\theta}_n, a), \tag{18.23}$$

where $\{\mathbf{S}_n\}$ is a sequence of symmetric matrices converging to a positive definite matrix $\mathbf{S}_0$. We define $\tilde{a}_n^0$ as a solution to this minimization problem.

First, we consider the Wald test based on $g(\hat{\theta}_n, \tilde{a}_n^0)$, i.e., we consider an appropriate quadratic form based on $g(\hat{\theta}_n, \tilde{a}_n^0)$ and a consistent estimator of a generalized inverse of the asymptotic variance covariance matrix of $\sqrt{n}g(\hat{\theta}_n, \tilde{a}_n^0)$.

Property 18.9: *If $\mathbf{S}_0 = \mathbf{\Omega}_0^{-1}$, then the Wald test based on $g(\hat{\theta}_n, \tilde{a}_n^0)$ is asymptotically equivalent to the tests ξ_n^W, ξ_n^L, ξ_n^H, and ξ_n^S defined in Sections 18.1.1 and 18.1.2.*

PROOF: Since $\mathbf{S}_0 = \mathbf{\Omega}_0^{-1}$ by assumption, it follows from Property 10.8 and the definition of $\mathbf{\Omega}_0$ that

$$\sqrt{n}(\hat{a}_n^0 - \tilde{a}_n^0) \# 0.$$

Hence

$$\sqrt{n}g(\hat{\theta}_n, \tilde{a}_n^0) \# \sqrt{n}g(\hat{\theta}_n, \hat{a}_n^0).$$

Thus the Wald test based on $\sqrt{n}g(\hat{\theta}_n, \tilde{a}_n^0)$ is asymptotically equivalent to the test based on $\sqrt{n}g(\hat{\theta}_n, \hat{a}_n^0)$, i.e., to ξ_n^W. The remaining asymptotic

equivalences follow from the asymptotic equivalences given in Properties 18.4 and 18.7. □

In the case where $\mathcal{I}_0 = \mathcal{J}_0$, we can again propose an asymptotically equivalent test that is equal to the optimal value of an objective function, which is here the function Φ_n.

Property 18.10: *If* $\mathbf{S}_0 = \mathbf{\Omega}_0^{-1}$ *and* $\mathcal{I}_0 = \mathcal{J}_0$, *then all the test statistics of Sections 18.1.1 and 18.1.2 are asymptotically equivalent to the optimal value of* Φ_n, *i.e., to* $\Phi_n(\tilde{a}_n^0)$ *under* H_0.

PROOF: This result is a direct consequence of Property 18.9 and the expression for ξ_n^W given in Property 18.6 when $\mathcal{I}_0 = \mathcal{J}_0$, namely

$$\xi_n^W = ng'(\hat{\theta}_n, \hat{a}_n^0)\hat{\mathbf{\Omega}}_n^{-1} g'(\hat{\theta}_n, \hat{a}_n^0).$$

□

When $\mathcal{I}_0 = \mathcal{J}_0$, note that

$$\begin{aligned} \mathbf{\Omega}_0 &= \frac{\partial g}{\partial \theta'} \mathcal{J}_0^{-1} \frac{\partial g'}{\partial \theta} \\ &= \frac{\partial g}{\partial \theta'} V_{as}\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right) \frac{\partial g'}{\partial \theta}. \end{aligned}$$

Hence, from Property 9.1, $\tilde{a}_n^0$ is the best asymptotic least squares estimator of a_0.

When $\mathcal{I}_0$ is not equal to $\mathcal{J}_0$, $\tilde{a}_n^0$ is no longer the best asymptotic least squares estimator of a_0. Hence one can think of using a more efficient estimator by considering the best asymptotic least squares estimator of a_0. As the next property shows, however, all test statistics based on asymptotic least squares estimators of a_0 are asymptotically equivalent.

Property 18.11: *Consider test statistics of the form*

$$ng(\hat{\theta}_n, \dot{a}_n^0)' V_{as}^{-}\left(\sqrt{n} g(\hat{\theta}_n, \dot{a}_n^0)\right) g'(\hat{\theta}_n, \dot{a}_n^0),$$

where $\dot{a}_n^0$ *is an asymptotic least squares (ALS) estimator of* a *obtained by solving (18.23) where* $\mathbf{S}_n$ *is an arbitrary matrix. Then all such test statistics are asymptotically equivalent. Moreover, they are asymptotically equivalent to the test statistics of Sections 18.1.1 and 18.1.2.*

PROOF: Let $\dot{a}_n^{(1)}$ be an ALS estimator of a associated with $\mathbf{S}_0^{(1)}$. We have

$$\sqrt{n} g(\hat{\theta}_n, \dot{a}_n^{(1)}) \# \frac{\partial g}{\partial \theta'} \sqrt{n}(\hat{\theta}_n - \theta_0) + \frac{\partial g}{\partial a'} \sqrt{n}(\dot{a}_n^{(1)} - a_0).$$

But from Theorem 9.2 we have

$$\sqrt{n}(\dot{a}_n^{(1)} - a_0) \# -\left(\frac{\partial g'}{\partial a}\mathbf{S}_0^{(1)}\frac{\partial g}{\partial a'}\right)^{-1}\frac{\partial g'}{\partial a}\mathbf{S}_0^{(1)}\frac{\partial g}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0).$$

Hence

$$\sqrt{n}g(\hat{\theta}_n, \dot{a}_n^{(1)}) \# (\mathbf{I} - \mathbf{M}^{(1)})\frac{\partial g}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0),$$

where

$$\mathbf{M}^{(1)} = \frac{\partial g}{\partial a'}\left(\frac{\partial g'}{\partial a}\mathbf{S}_0^{(1)}\frac{\partial g}{\partial a'}\right)^{-1}\frac{\partial g'}{\partial a}\mathbf{S}_0^{(1)}.$$

Similarly, if $\dot{a}_n^{(2)}$ is another ALS estimator of a associated with $\mathbf{S}_0^{(2)}$, we obtain

$$\sqrt{n}g(\hat{\theta}_n, \dot{a}_n^{(2)}) \# (\mathbf{I} - \mathbf{M}^{(2)})\frac{\partial g}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0),$$

where

$$\mathbf{M}^{(2)} = \frac{\partial g}{\partial a'}\left(\frac{\partial g'}{\partial a}\mathbf{S}_0^{(2)}\frac{\partial g}{\partial a'}\right)^{-1}\mathbf{S}_0^{(2)}.$$

Now, there is a bijective relationship between

$$\mathbf{A}_1 = (\mathbf{I} - \mathbf{M}^{(1)})\frac{\partial g}{\partial \theta'}\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$$

and

$$\mathbf{A}_2 = (\mathbf{I} - \mathbf{M}^{(2)})\frac{\partial g}{\partial \theta'}\sqrt{n}\left(\hat{\theta}_n - \theta_0\right).$$

This follows from

$$(\mathbf{I} - \mathbf{M}^{(2)})\mathbf{A}_1 = \mathbf{A}_2 \quad \text{and} \quad (\mathbf{I} - \mathbf{M}^{(1)})\mathbf{A}_2 = \mathbf{A}_1.$$

Hence the various test statistics defined in the statement of Property 18.11 are asymptotically equivalent. In addition, they are asymptotically equivalent to the test statistics defined earlier because of Property 18.9. □

In particular, Property 18.11 can be used to extend Property 18.10. That is, there exists an asymptotically equivalent test statistic based on the optimal value of an objective function even when $\mathcal{I}_0 \neq \mathcal{J}_0$.

Property 18.12: *Let $\bar{a}_n^0$ be the ALS estimator of a_0 obtained by minimizing*

$$\bar{\Phi}_n(a) = ng'(\hat{\theta}_n, a)\bar{\mathbf{S}}_n g(\hat{\theta}_n, a),$$

where $\bar{\mathbf{S}}_n$ *converges to*

$$\begin{aligned}\bar{\mathbf{S}}_0 &= \left(\frac{\partial g}{\partial \theta'} V_{as}\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right)\frac{\partial g'}{\partial \theta}\right)^{-1} \\ &= \left(\frac{\partial g}{\partial \theta}\mathcal{J}_0^{-1}\mathcal{I}_0\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}\right)^{-1}.\end{aligned}$$

Then the test statistic $\bar{\Phi}_n(\bar{a}_n^0)$ *is asymptotically equivalent to the test statistics introduced previously under* H_0.

PROOF: From Property 18.11 it follows that the test statistic

$$ng'(\hat{\theta}_n, \bar{a}_n^0)\hat{V}_{as}^{-}\left(\sqrt{n}g(\hat{\theta}_n, \bar{a}_n^0)\right)g(\hat{\theta}_n, \bar{a}_n^0)$$

is asymptotically equivalent to the test statistics introduced previously. It remains to show that $\bar{\mathbf{S}}_0$ is a generalized inverse of $V_{as}\left(\sqrt{n}g(\hat{\theta}_n, \bar{a}_{0n})\right)$. But, from the proof of Property 18.11 we know that

$$\begin{aligned}V_{as}\left(\sqrt{n}g(\hat{\theta}_n, \bar{a}_{0n})\right) &= (\mathbf{I} - \bar{\mathbf{M}})\frac{\partial g}{\partial \theta'}\mathcal{J}_0^{-1}\mathcal{I}_0\mathcal{J}_0^{-1}\frac{\partial g'}{\partial \theta}(\mathbf{I} - \bar{\mathbf{M}}) \\ &= (\mathbf{I} - \bar{\mathbf{M}})\bar{\mathbf{S}}_0^{-1}(\mathbf{I} - \bar{\mathbf{M}}),\end{aligned}$$

where

$$\bar{\mathbf{M}} = \frac{\partial g}{\partial a'}\left(\frac{\partial g'}{\partial a}\bar{\mathbf{S}}_0\frac{\partial g}{\partial a'}\right)^{-1}\frac{\partial g'}{\partial a}\bar{\mathbf{S}}_0.$$

Then the desired result follows immediately from the properties of the projection matrices established in the proof of Property 18.6. □

In the general case where $\mathcal{I}_0 \neq \mathcal{J}_0$, Property 18.12 provides a particularly convenient test since its test statistic is obtained directly from the optimal value of the objective function $\bar{\Phi}_n(a)$. In addition, this testing procedure does not require the computation of generalized inverses and hence avoids the problem of discontinuity mentioned earlier.

18.1.4 An Example: Tests Based on Least Squares

We consider the second-order econometric model defined by

$$\begin{cases} Y_i = h(X_i, b) + u_i, \\ E(u_i \mid X) = 0, V(u_i \mid X) = \gamma\omega^2(X_i, b). \end{cases} \tag{18.24}$$

We want to test the null hypothesis under the implicit form

$$H_0 : g(b) = 0. \tag{18.25}$$

a) Ordinary Least Squares

The unconstrained and constrained ordinary least squares (OLS) estimators of b are obtained by maximizing the objective function

$$L_n(b) = -\frac{1}{2}\sum_{i=1}^{n}(Y_i - h(x_i, b))^2$$

possibly subject to the constraints (18.25). We let $\hat{b}_n$ and $\hat{b}_n^0$ denote these estimators, respectively.

These OLS estimators can be viewed as pseudo maximum likelihood estimators associated with a normal distribution with known variance. From equations (8.11) and (8.12), the matrices $\mathcal{I}_0$ and $\mathcal{J}_0$ are

$$\mathcal{J}_0 = E_X\left(\frac{\partial h'(X, b_0)}{\partial b}\frac{\partial h(X, b_0)}{\partial b'}\right)$$

and

$$\mathcal{I}_0 = \gamma_0 E_X\left(\frac{\partial h'(X, b_0)}{\partial b}\frac{\partial h(X, b_0)}{\partial b'}\omega^2(X, b_0)\right),$$

where b_0 and γ_0 are the true values of b and γ, respectively. In general, these two matrices are not equal.

These matrices can be estimated by

$$\hat{\mathcal{J}}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial h'(x_i, \hat{b}_n)}{\partial b}\frac{\partial h(x_i, \hat{b}_n)}{\partial b'}$$

and

$$\hat{\mathcal{I}}_n = \frac{1}{n}\hat{\gamma}_n\sum_{i=1}^{n}\frac{\partial h'(x_i, \hat{b}_n)}{\partial b}\frac{\partial h(x_i, \hat{b}_n)}{\partial b'}\omega^2(x_i, \hat{b}_n),$$

where

$$\hat{\gamma}_n = \sum_{i=1}^{n}\frac{\left(y_i - h(x_i, \hat{b}_n)\right)^2}{\omega^2(x_i, \hat{b}_n)}.$$

Alternatively, the matrices $\mathcal{I}_0$ and $\mathcal{J}_0$ can be estimated by $\hat{\mathcal{I}}_n^0$ and $\hat{\mathcal{J}}_n^0$, which are obtained by replacing $\hat{b}_n$ by $\hat{b}_n^0$ in the preceding expressions. Let $\hat{\gamma}_n^0$ correspond to $\hat{b}_n^0$.

We can now determine the various test statistics defined in the preceding sections. For instance, noting that $\mathbf{M} = 0$, the Lagrange multiplier test statistic defined in Property 18.2 becomes

$$\xi_n^L = \frac{1}{n}\hat{\lambda}_n'\hat{\Omega}_n^0\left(\frac{\partial g(\hat{b}_n^0)}{\partial b'}(\hat{\mathcal{J}}_n^0)^{-1}\hat{\mathcal{I}}_n^0(\hat{\mathcal{J}}_n^0)^{-1}\frac{\partial g'(\hat{b}_n^0)}{\partial b}\right)^{-1}\hat{\Omega}_n^0\hat{\lambda}_n,$$

where

$$\hat{\Omega}_n^0 = \frac{\partial g(\hat{b}_n^0)}{\partial b'}(\hat{\mathcal{J}}_n^0)^{-1}\frac{\partial g'(\hat{b}_n^0)}{\partial b}.$$

In particular, for testing the hypothesis $H_0 : b = b_0$, we have

$$\frac{\partial g}{\partial b'} = \mathbf{I}, \quad \Omega = \mathcal{J}^{-1}, \quad \text{and} \quad \hat{\lambda}_n = \frac{\partial L_n(\hat{b}_n^0)}{\partial b'}.$$

Hence the test statistic ξ_n^L becomes

$$\begin{aligned}
\xi_n^L &= \frac{1}{n}\frac{\partial L_n(\hat{b}_n^0)}{\partial b'}\left(\hat{\mathcal{J}}_n^0\right)^{-1}\hat{\mathcal{J}}_n^0\left(\hat{\mathcal{I}}_n^0\right)^{-1}\hat{\mathcal{J}}_n^0\left(\hat{\mathcal{J}}_n^0\right)^{-1}\frac{\partial L_n(\hat{b}_n^0)}{\partial b} \\
&= \frac{1}{n}\frac{\partial L_n(\hat{b}_n^0)}{\partial b'}\left(\hat{\mathcal{I}}_n^0\right)^{-1}\frac{\partial L_n(\hat{b}_n^0)}{\partial b} \\
&= \frac{1}{n\hat{\gamma}_n^0}\sum_{i=1}^n\left(y_i - h(x_i,\hat{b}_n^0)\right)\frac{\partial h(x_i,\hat{b}_n^0)}{\partial b'} \\
&\quad \times\left(\sum_{i=1}^n\frac{\partial h'(x_i,\hat{b}_n^0)}{\partial b}\frac{\partial h(x_i,\hat{b}_n^0)}{\partial b'}\omega^2(x_i,\hat{b}_n^0)\right)^{-1} \\
&\quad \times\sum_{i=1}^n\left(y_i - h(x_i,\hat{b}_n^0)\right)\frac{\partial h'(x_i,\hat{b}_n^0)}{\partial b}.
\end{aligned}$$

Note that this test statistic is based on the scalar product of the residuals $\hat{u}_i^0 = y_i - h(x_i,\hat{b}_n^0)$ under H_0 and the "explanatory variables" $\partial h(x_i;\hat{b}_n^0)/\partial b$.

b) Generalized Least Squares

Let $\tilde{b}_n$ be a consistent estimator of b. For instance, $\tilde{b}_n$ can be the unconstrained OLS estimator of b. The generalized least squares (GLS) estimator is based on the objective function

$$L_n(b) = -\frac{1}{2\tilde{\gamma}_n}\sum_{i=1}^n\frac{(y_i - h(x_i,b))^2}{\omega^2(x_i,\tilde{b}_n)},$$

where $\tilde{\gamma}_n$ is a consistent estimator of γ_0. We let $\hat{b}_n$ and $\hat{b}_n^0$ denote the unconstrained and constrained GLS estimators of b.

The GLS estimation method can be viewed as a quasi generalized pseudo maximum likelihood estimation method associated with a normal distribution with unknown variance. We know that $\mathcal{I}_0 = \mathcal{J}_0$ in this case. Hence we have

$$\mathcal{I}_0 = \mathcal{J}_0 = E_X \left(\frac{1}{\gamma_0 \omega^2(X, b_0)} \frac{\partial h'(X, b_0)}{\partial b} \frac{\partial h(X, b_0)}{\partial b'} \right).$$

We can apply for instance Property 18.5. Hence a test statistic is

$$\sum_{i=1}^{n} \frac{\left(y_i - h(x_i, \hat{b}_n^0)\right)^2}{\tilde{\gamma}_n \omega^2(x_i, \tilde{b}_n)} - \sum_{i=1}^{n} \frac{\left(y_i - h(x_i, \hat{b}_n)\right)^2}{\tilde{\gamma}_n \omega^2(x_i, \tilde{b}_n)}.$$

18.2 Specification Tests Based on Asymptotic Least Squares

18.2.1 General Principle

In this section we no longer require that the unconstrained estimator $\hat{\theta}_n$ of θ is obtained from the maximization of an objective function. Instead we only require that this estimator is consistent and asymptotically normal, i.e.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathbf{Q}_0), \tag{18.26}$$

where $\mathbf{Q}_0$ is a some nonsingular matrix.

As before, we want to test the null hypothesis under the mixed form given in (18.5), i.e.

$$H_0 : \{\theta : \exists a \in A \subset \mathbb{R}^q, g(\theta, a) = 0\}, \tag{18.27}$$

where g takes its values in $\mathbb{R}^r$. It is assumed that the rank conditions (18.6) are satisfied, namely

$$\operatorname{rank}\left(\frac{\partial g}{\partial \theta'}\right) = r \quad \text{and} \quad \operatorname{rank}\left(\frac{\partial g}{\partial a'}\right) = q, \tag{18.28}$$

More generally, we assume that the assumptions of Property 9.1 are satisfied.

The test that we shall develop can be viewed as a specification test in the sense that it assesses the validity of the specification $g(\theta, a) = 0$, which defines the parameter a.

Specifically, we can estimate a under H_0 by the asymptotic least squares method, that minimizes with respect to a

$$g(\hat{\theta}_n, a)' \mathbf{S}_n^* g(\hat{\theta}_n, a), \tag{18.29}$$

where $\{\mathbf{S}_n^*\}$ is a sequence converging to

$$\mathbf{S}_0^* = \left(\frac{\partial g}{\partial \theta'} \mathbf{Q}_0 \frac{\partial g'}{\partial \theta} \right)^{-1}$$

evaluated at θ_0 and a_0, which are the true values of θ and a, respectively. The matrix $\mathbf{S}_0^{*-1}$ is the asymptotic variance covariance matrix of $\sqrt{n} g(\hat{\theta}_n, a_0)$. For instance, $\mathbf{S}_n^*$ can be obtained by replacing $\mathbf{Q}_0$ by a consistent estimator $\hat{\mathbf{Q}}_n$, θ_0 by $\hat{\theta}_n$ and a_0 by $\tilde{a}_n$, where $\tilde{a}_n$ is an ALS estimator of a associated with an arbitrary sequence of matrices $\mathbf{S}_n$, for instance $\mathbf{S}_n = \mathbf{I}$.

The null hypothesis (18.27) can be tested using the next property.

Property 18.13: *Let $\hat{a}_n$ be the ALS estimator of a obtained by minimizing (18.29). Under the null hypothesis H_0 defined in (18.27), the test statistic $\xi_n^A = n g'(\hat{\theta}_n, \hat{a}_n) \mathbf{S}_n^* g(\hat{\theta}_n, \hat{a}_n)$ is asymptotically distributed as a $\chi^2(r-q)$. Moreover, the test with critical region $\{\xi_n^A \geq \chi^2_{1-\alpha}(r-q)\}$ has an asymptotic level equal to α and is consistent.*

PROOF: The asymptotic distribution of ξ_n^A can be obtained as in the proof of Property 18.12 by replacing $\mathcal{J}_0^{-1} \mathcal{I}_0 \mathcal{J}_0^{-1}$ by $\mathbf{Q}_0$. On the other hand, if the null hypothesis H_0 does not hold, then $\hat{a}_n$ converges to a value a_∞ such that $g(\theta_0, a_\infty) \neq 0$. Hence ξ_n^A converges to $+\infty$ and the test is consistent. □

Since $(\mathbf{S}_0^*)^{-1}$ is the asymptotic variance covariance matrix of the vector $\sqrt{n} g(\hat{\theta}_n, a_0)$, then $n g(\hat{\theta}_n, a_0) S_n^* g(\hat{\theta}_n, a_0)$ is distributed asymptotically as a $\chi^2(r)$ under H_0. As a result, by replacing a_0 by $\hat{a}_n$ in this expression, Property 18.13 shows that we obtain a test statistic that is also χ^2 distributed but with a number of degrees of freedom which is reduced by the number q of components of a.

18.2.2 Unknown Linear Constraints on Functions of the Parameter θ

Suppose that $g(\theta, a)$ is linear in a so that the null hypothesis H_0 becomes

$$H_0 : \{\exists a : h(\theta) = \mathbf{H}(\theta)a\}, \tag{18.30}$$

where $h(\cdot)$ is a known vector function of dimension r and $\mathbf{H}(\cdot)$ is a known $r \times q$ matrix function. Thus the null hypothesis H_0 consists of unknown linear constraints on known functions of θ.

The rank conditons become: (i) The matrix

$$\mathbf{M}(\theta, a) = \frac{\partial h(\theta)}{\partial \theta'} - \sum_{i=1}^{q} \frac{\partial h_i(\theta)}{\partial \theta'} a_i$$

is of rank r, where $h_i(\theta)$ is the i-th column of the matrix $\mathbf{H}(\theta)$ and a_i is the i-th component of a, and (ii) the matrix $\mathbf{H}(\theta)$ is of rank q.

When the null hypothesis is of the form (18.30), the test proposed in Property 18.13 becomes particularly simple. Namely, the testing procedure consists of:

(i) Regressing $h(\hat{\theta}_n)$ on the columns of $\mathbf{H}(\hat{\theta}_n)$ by the OLS method. This provides $\tilde{a}_n$.

(ii) Computing

$$\hat{\mathbf{M}}_n = \mathbf{M}(\hat{\theta}_n, \tilde{a}_n) \quad \text{and} \quad \mathbf{S}_n^{*-1} = \hat{\mathbf{M}}_n \hat{\mathbf{Q}}_n \hat{\mathbf{M}}_n'.$$

(iii) Regressing $h(\hat{\theta}_n)$ on the columns of $\mathbf{H}(\hat{\theta}_n)$ by the GLS method based upon $\mathbf{S}_n^{*-1}$. This provides $\hat{a}_n$. Moreover, n times the optimal value of the objective function gives the test statistic, i.e.

$$\xi_n^A = n \left(h(\hat{\theta}_n) - \mathbf{H}(\hat{\theta}_n)\hat{a}_n\right)' \mathbf{S}_n^* \left(h(\hat{\theta}_n) - \mathbf{H}(\hat{\theta}_n)\hat{a}_n\right)$$

The critical region of the test is

$$W = \left\{\xi_n^A \geq \chi_{1-\alpha}^2(r-q)\right\}.$$

18.2.3 Examples

a) Rational Lags

Consider a linear model of the form

$$y_t = \beta z_t + \theta_1 x_t + \theta_2 x_{t-1} + \ldots + \theta_p x_{t-p+1} + u_t, \qquad (18.31)$$

for $t = 1, \ldots, T$, where the u_t's are uncorrelated with zero means and common variance σ^2. The variables z_t and x_t are exogenous, and z_t is possibly a vector while x_t is a scalar.

Estimation by ordinary least squares gives a consistent and asymptotically normal estimator. In particular, let $\hat{\theta}_n$ be the (unconstrained) OLS estimator of θ, where $\theta' = (\theta_1, \ldots, \theta_p)$.

Let L denote the lag operator, i.e., $Lx_t = x_{t-1}$. One wants to test the hypothesis that $\Theta(L) = \theta_1 + \theta_2 L + \ldots + \theta_p L^{p-1}$ is equal to a rational polynomial fraction $C(L)/D(L)$ whose expansion is truncated at the order $p-1$ and where the degree of C is m_0 and the degree of D is m_1. By convention we use the normalization $D(0) = 1$.

To simplify, consider the case where $m_0 = 0$ and $m_1 = 1$. Then the null hypothesis is

$$H_0 : \left\{ \exists d, c : \Theta(L) = \frac{c}{1 - dL} \quad \text{truncated at} \quad p-1 \right\}.$$

The null hypothesis is a *Koyck type hypothesis.*

This hypothesis can be written as

$$\exists d : \begin{cases} \theta_2 - d\,\theta_1 = 0, \\ \vdots \\ \theta_p - d\,\theta_{p-1} = 0. \end{cases}$$

Thus the null hypothesis is of the form (18.30). Hence the test statistic for testing H_0 is readily obtained by applying the general procedure described in Section 18.2.2. Namely,

(i) Regress by ordinary least squares

$$\begin{pmatrix} \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_p \end{pmatrix} \quad \text{on} \quad \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_{p-1} \end{pmatrix}.$$

This provides $\tilde{d}_T$.

(ii) Compute the variance covariance matrix of

$$\begin{pmatrix} \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_p \end{pmatrix} - d \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_{p-1} \end{pmatrix} \equiv \mathbf{S}^*(d, \sigma^2)^{-1}.$$

Then estimate it by $\mathbf{S}_T^{*-1} = \mathbf{S}^*(\tilde{d}_T, \hat{\sigma}_T^2)^{-1}$, where $\hat{\sigma}_T^2$ is a consistent estimator of σ^2 obtained by OLS on (18.31).

(iii) Repeat (i) but using GLS associated with $\mathbf{S}_T^{*-1}$. Then compute T times the optimal value of the GLS objective function. This gives the test statistic ξ_T^A.

The critical region of the test is $\{\xi_T^A \geq \chi^2_{1-\alpha}(p-2)\}$ since $r = p-1$ and $q = 1$ here. Such a test can be easily generalized to the case where the degrees of C and D are arbitrary (see Exercise 18.3).

b) Tests of Validity of Instrumental Variables: The Linear Case

Consider one equation of a simultaneous equation model. Let this equation be

$$y_t = x_t b + u_t, \tag{18.32}$$

where x_t is a row vector of dimension k possibly including some endogenous variables and u_t is a random error with common variance σ^2.

One wants to test the hypothesis that the row vector z_t of dimension ℓ is a valid vector of instrumental variables, i.e., a vector satisfying

$$p\lim_T \frac{1}{T} \sum_{t=1}^T z_t u_t = 0.$$

Such a hypothesis can also be written as

$$H_0 : \left\{ \exists b : p\lim_T \frac{1}{T} \sum_{t=1}^T z_t' y_t = \left(p\lim_T \frac{1}{T} \sum_{t=1}^T z_t' x_t \right) b \right\}. \tag{18.33}$$

Thus the null hypothesis can be written as an unknown linear constraint on the parameters $p\lim_T \frac{1}{T} \sum_{t=1}^T z_t' y_t$ and $p\lim_T \frac{1}{T} \sum_{t=1}^T z_t' x_t$. We stack these parameters as follows

$$\theta = \left(p\lim_T \frac{1}{T} \sum_{t=1}^T z_t y_t, p\lim_T \frac{1}{T} \sum_{t=1}^T z_t x_{1t}, \ldots, p\lim_T \frac{1}{T} \sum_{t=1}^T z_t x_{kt} \right)'.$$

A consistent estimator of θ is

$$\hat{\theta}_T = \frac{1}{T}\left(\sum_{t=1}^{T} z_t y_t, \sum_{t=1}^{T} z_t x_{1t}, \ldots, \sum_{t=1}^{T} z_t x_{kt}\right)'.$$

To satisfy the conditions underlying the general method of Section 18.2.2, it is assumed that the matrix $p\lim_T (1/T)\sum_{t=1}^{T} z_t' x_t$ is of rank k. This matrix plays the same role as $\mathbf{H}(\theta)$. This assumption implies that $\ell \geq k$. It is also assumed that the asymptotic variance covariance matrix of $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is nonsingular. In particular, this excludes the case where the components of z_t are linearly dependent. Lastly, it is easy to verify that the matrix $\mathbf{M}(\theta, a)$ is of full rank ℓ as required.

The asymptotic variance covariance matrix $\mathbf{S}_0^{*-1}$ of

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} z_t' y_t - \frac{1}{T}\sum_{t=1}^{T} z_t' x_t b\right),$$

i.e., of $\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} z_t' u_t\right)$, is equal to $\sigma^2 p\lim_T \frac{1}{T}\sum_{t=1}^{T} z_t z_t'$ under general conditions. A sequence of matrices converging to this matrix is given by $\sigma^2 \frac{1}{T}\sum_{t=1}^{T} z_t z_t'$, which is known up to a multiplicative constant. Thus we can apply directly the asymptotic least squares method (see Section 9.4.5) and minimize with respect to b

$$\frac{1}{\sigma^2}\left(\frac{1}{T}\sum_{t=1}^{T} z_t' y_t - \frac{1}{T}\sum_{t=1}^{T} z_t' x_t b\right)' \times \left(\frac{1}{T}\sum_{t=1}^{T} z_t' z_t\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} z_t' y_t - \frac{1}{T}\sum_{t=1}^{T} z_t' x_t b\right).$$

Clearly the constant $1/\sigma^2$ does not change the solution to the above minimization problem. This solution is equal to the two-stage least squares estimator

$$\hat{b} = \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y},$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the matrices whose t-th rows are x_t and z_t respectively.

The test statistic obtained by applying the general procedure of Section 18.2.2 is equal to T times the optimal value of the above objective

function. It is, however, necessary to take into account the presence of $1/\sigma^2$. A natural estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{T} \|\hat{u}\|^2,$$

where $\hat{u} = y - \mathbf{X}\hat{b}$. Thus the test statistic is

$$\xi_T = \frac{1}{\hat{\sigma}^2} \hat{u}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{u}.$$

The critical region of the test at asymptotic level α is

$$W = \left\{\xi_T \geq \chi^2_{1-\alpha}(l-k)\right\}.$$

Note that ξ_T can be written as $T \times \|\mathbf{P}_Z\hat{u}\|^2/\|\hat{u}\|^2$, where $\mathbf{P}_Z$ is the orthogonal projection onto the subspace generated by the columns of $\mathbf{Z}$. That is, ξ_T is equal to TR^2, where R^2 is the (uncentered) coefficient of determination in the regression of $\hat{u}$ on the columns of $\mathbf{Z}$.

c) Tests of Overidentifying Restrictions

Suppose that the process generating the vector (X_t, Y_t) is a bivariate autoregressive process of order p, i.e., a process defined by

$$\left\{ \begin{array}{rcl} X_t &=& \phi_{xx}(L)X_t + \phi_{xy}(L)Y_t + \epsilon_t, \\ Y_t &=& \phi_{yx}(L)X_t + \phi_{yy}(L)Y_t + \eta_t, \end{array} \right. \tag{18.34}$$

where L is the lag operator and ϕ_{xy}, ϕ_{xx}, ϕ_{yx}, and ϕ_{yy} are polynomials of degree p with no constant terms. It is also assumed that the process for (ϵ_t, η_t) is white noise with an arbitrary variance covariance matrix. Lastly, it is assumed that the process is stationary, i.e., that the matrix

$$\begin{pmatrix} 1 - \Phi_{xx}(L) & -\Phi_{xy}(L) \\ -\Phi_{yx}(L) & 1 - \Phi_{yy}(L) \end{pmatrix}$$

has a determinant whose roots are all larger than one in modulus.

The coefficients of the above polynomials can be estimated by the maximum likelihood method (conditionally upon the first p observations) assuming that the process is Gaussian. Alternatively, such an estimator is equal to the ordinary least squares estimator obtained from each equation taken separately. The latter estimator has the same asymptotic properties but does not require the normality assumption.

One wants to test that the model is actually the simultaneous equation model

$$\begin{cases} X_t &= \alpha Y_t + \beta X_{t-1} + u_t, \\ Y_t &= \gamma X_t + \delta Y_{t-2} + \nu_t, \end{cases} \tag{18.35}$$

where the variance covariance matrix of the vector (u_t, v_t) is unconstrained.

The null hypothesis (18.35) implies some constraints on the system (18.34), which can be written as

$$\exists\, \alpha, \beta, \gamma, \delta : \begin{pmatrix} 1 & -\alpha \\ -\gamma & 1 \end{pmatrix} \begin{pmatrix} \Phi_{xx}(L) & \Phi_{xy}(L) \\ \Phi_{yx}(L) & \Phi_{yy}(L) \end{pmatrix} = \begin{pmatrix} \beta L & 0 \\ 0 & \delta L^2 \end{pmatrix}. \tag{18.36}$$

In terms of the coefficients of the polynomials in (18.34), we obtain a null hypothesis under mixed form that is linear in the parameters $(\alpha, \beta, \gamma, \delta)$. These play the role of the auxiliary parameter a in the theory presented earlier. Thus the testing procedure of Section 18.2.2 applies directly.

This procedure can be easily extended to null hypotheses that are more complex than (18.35). See Exercise 18.4.

18.3 Specification Tests Based on Generalized Methods of Moments

18.3.1 General Result

From Section 9.5, a generalized method of moments (GMM) is based on estimating equations of the form

$$E_0 h(Y, X, \theta_0) = 0, \tag{18.37}$$

where h is a H-dimensional vector function, θ is a p-dimensional parameter vector whose true value is θ_0, and E_0 is the expectation operator with respect to the true joint distribution of (Y, X). Some regularity assumptions are imposed (see Section 9.5). In particular, assumption H12 requires that h is differentiable with respect to θ and that the rank of the $H \times p$ matrix $\partial h / \partial \theta'$ is equal to p so that $H \geq p$.

A generalized method of moment estimator is obtained by minimizing

$$\left(\sum_{i=1}^{n} h(Y_i, X_i, \theta)\right)' \mathbf{S}_n \left(\sum_{i=1}^{n} h(Y_i, X_i, \theta)\right),$$

where $\{\mathbf{S}_n\}$ is a sequence of matrices converging to a matrix $\mathbf{S}_0$. From Property 9.10 we know that the best GMM estimator θ_n^* is obtained when the limit matrix is $\mathbf{S}_0^* = (V_0 h(Y, X, \theta_0))^{-1}$.

In this section we shall study how to test that (18.37) is actually satisfied. This says that there exists a unique value θ_0 satisfying (18.37) for each possible distribution of (Y, X). Thus, intuitively, such a specification is restrictive only when $H > p$. Hence a test can be based on the following property.

Property 18.14: *Let $\hat{V}_h$ be a consistent estimator of $V_0 h(Y, X, \theta_0)$. Assume that the hypotheses of Section 9.5 hold. Then the statistic*

$$\xi_n^G = \frac{1}{n} \sum_{i=1}^{n} h'(Y_i, X_i, \theta_n^*)(\hat{V}h)^{-1} \sum_{i=1}^{n} h'(Y_i, X_i, \theta_n^*)$$

is asymptotically distributed as a $\chi^2(H - p)$ under H_0 as defined by (18.37). The test with critical region $\{\xi_n^G \geq \chi^2_{1-\alpha}(H - p)\}$ has asymptotic level α and is consistent.

PROOF: We have

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(Y_i, X_i, \theta_n^*) \\ &\quad \# \quad \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} h(Y_i, X_i, \theta_0) + \sum_{i=1}^{n} \frac{\partial h(Y_i, X_i, \theta_0)}{\partial \theta'} (\theta_n^* - \theta_0) \right) \\ &\quad \# \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(Y_i, X_i, \theta_0) + E_0 \frac{\partial h(Y, X, \theta_0)}{\partial \theta'} \sqrt{n}(\theta_n^* - \theta_0). \end{aligned}$$

On the other hand, the proof of Property 9.9 implies

$$\begin{aligned} \sqrt{n}(\theta_n^* - \theta_0) \quad \# \quad & -\left(E_0 \frac{\partial h'(Y, X, \theta_0)}{\partial \theta} (V_0 h)^{-1} E_0 \frac{\partial h'(Y, X, \theta_0)}{\partial \theta} \right)^{-1} \\ & \times E_0 \frac{\partial h'(Y, X, \theta_0)}{\partial \theta} (V_0 h)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(Y_i, X_i, \theta_0). \end{aligned} \tag{18.38}$$

Hence, simplifying notation, we obtain

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(Y_i, X_i, \theta_n^*) \\
&\quad \# \quad \left(\mathbf{I} - E_0\frac{\partial h}{\partial \theta'}\left(E_0\frac{\partial h}{\partial \theta'}(V_0 h)^{-1}E_0\frac{\partial h}{\partial \theta'}\right)^{-1} E_0\frac{\partial h'}{\partial \theta}(V_0 h)^{-1}\right) \\
&\qquad \times\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(Y_i, X_i, \theta_0) \\
&\quad = \quad (\mathbf{I} - \mathbf{M}_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(Y_i, X_i, \theta_0),
\end{aligned}$$

where $\mathbf{M}_0$ is the orthogonal projection matrix onto the subspace generated by the columns of $E_0(\partial h/\partial\theta')$ with respect to the scalar product defined by $(V_0 h)^{-1}$.

Thus the asymptotic distribution of $(1/\sqrt{n})\sum_{i=1}^{n} h(Y_i, X_i, \theta_n^*)$ is the singular multivariate normal distribution $N(0, (\mathbf{I} - \mathbf{M}_0)V_0 h(\mathbf{I} - \mathbf{M}_0'))$. Since $(V_0 h)^{-1}$ is a generalized inverse of $(\mathbf{I} - \mathbf{M}_0)V_0 h(\mathbf{I} - \mathbf{M}_0')$ (see the proof of Property 18.6), it follows that the asymptotic distribution of ξ_n^G is a χ^2 distribution with the number of degrees of freedom equal to the rank of $(\mathbf{I} - \mathbf{M}_0)V_0 h(\mathbf{I} - \mathbf{M}_0')$, which is $H - p$.

If there does not exist a value for the parameter θ such that specification (18.37) holds, then θ_n^* converges to θ_∞ (say) satisfying $E_0 h(Y, X, \theta_\infty) \neq 0$. It follows that ξ_n^G converges to $+\infty$. Hence the test is consistent. □

Note that the quadratic form

$$\frac{1}{n}\sum_{i=1}^{n} h'(Y_i, X_i, \theta_0)(\hat{V}h)^{-1}\sum_{i=1}^{n} h(Y_i, X_i, \theta_0)$$

is asymptotically distributed as a χ^2 distribution with H degrees of freedom under H_0. Hence Property 18.14 shows that replacing θ_0 by θ_n^* in the preceding quadratic form produces a statistic whose asymptotic distribution is still χ^2 but with a number of degrees of freedom smaller by p, which is the dimension of the parameter vector θ.

18.3.2 Examples

a) Tests of Validity of Instrumental Variables: The Nonlinear Case

Consider the nonlinear model

$$y_t = m(x_t, b) + u_t, \qquad t = 1, \ldots, T, \tag{18.39}$$

where x_t is a row vector possibly containing lagged endogenous variables, the u_t's are independent errors with zero mean and common variance σ^2, and b is a p-dimensional parameter vector.

The hypothesis of interest is that a row vector z_t of size $\ell > p$ is a valid vector of instrumental variables, i.e., is uncorrelated with the errors

$$E_0 z' u = 0.$$

Such a hypothesis can also be written as

$$H_0 : \{\exists b : E_0 z'(y - m(x, b)) = 0\}. \tag{18.40}$$

From Section 9.5.4 we know that the generalized method of moment estimator based on the preceding estimating constraints is the two-stage least squares estimator $\hat{b}$. Thus the statistic ξ_n^G becomes

$$\xi_n^G = \frac{1}{\hat{\sigma}^2} \hat{u}' \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \hat{u},$$

where $\hat{u}$ is the vector of residuals whose t-th component is $\hat{u}_t = y_t - m(x_t, \hat{b})$ and $\hat{\sigma}^2 = \frac{1}{T}\|\hat{u}\|^2$. Therefore a test of validity of the instrumental variables z_t has the critical region

$$\left\{\xi_n^G \geq \chi^2_{1-\alpha}(l - p)\right\}.$$

b) Specification Tests of a Parametric Conditional Model

Consider a parametric conditional model whose densities are $f(y \mid x; \theta)$, where θ is a p-dimensional vector.

We know that the score vector is centered conditionally upon X. Hence

$$E_0 \frac{\partial \log f(Y \mid X; \theta_0)}{\partial \theta} = 0.$$

Let $\mathbf{G}(X)$ be an arbitrary $r \times p$ matrix. Then we have

$$\begin{aligned} & E_0\left(\mathbf{G}(X)\frac{\partial \log f(Y \mid X;\theta_0)}{\partial\theta}\right) \\ &= E_0\left(\mathbf{G}(X)E_{\theta_0}\frac{\partial \log f(Y \mid X;\theta_0)}{\partial\theta} \mid X\right) \\ &= 0. \end{aligned}$$

Thus we can consider the estimating equations

$$E_0 h(Y,X,\theta) = 0, \quad \text{with} \quad h(Y,X,\theta) = \begin{pmatrix} \partial \log f(Y \mid X;\theta_0)/\partial\theta \\ \mathbf{G}(X)\partial \log f(Y \mid X;\theta_0)/\partial\theta \end{pmatrix}.$$

Here the number of estimating equations is $H = p + r$.

The GMM estimator θ_n^* based on h must be at least as preferred to the GMM estimator based on the first p components of h only. But the latter estimator is the maximum likelihood estimator of θ. Given the optimal asymptotic property of the maximum likelihood estimator (see Chapter 7), we can expect that θ_n^* is asymptotically equivalent to the maximum likelihood estimator. This can be formally established as follows.

From Property 9.9, the asymptotic variance covariance matrix of θ_n^* is

$$\mathbf{\Sigma}_0 = \left(E_0\frac{\partial h'(Y,X,\theta_0)}{\partial\theta}\left(V_0 h(Y,X,\theta_0)\right)^{-1}E_0\frac{\partial h(Y,X,\theta_0)}{\partial\theta'}\right)^{-1}.$$

But we have

$$\frac{\partial h}{\partial\theta'} = \begin{pmatrix} \mathbf{I} \\ \mathbf{G}(X) \end{pmatrix}\frac{\partial^2 \log f(Y,X,\theta)}{\partial\theta\partial\theta'}$$

and

$$E_0\frac{\partial h}{\partial\theta'} = -E_0\left(\begin{pmatrix} \mathbf{I} \\ \mathbf{G}(X) \end{pmatrix}\mathcal{I}_c(X,\theta_0)\right) = -\begin{pmatrix} \mathcal{I}(\theta_0) \\ E_0(\mathbf{G}(X)\mathcal{I}_c(X,\theta_0)) \end{pmatrix},$$

where $\mathcal{I}_c(X,\theta)$ is the Fisher information matrix conditional upon X. Similarly, we have

$$\begin{aligned} V_0 h(Y,X,\theta_0) &= E_0(h(Y,X,\theta_0)h'(Y,X,\theta_0)) \\ &= \begin{pmatrix} \mathcal{I}(\theta_0) & E_0(\mathbf{G}(X)\mathcal{I}_c(X,\theta_0))' \\ E_0(\mathbf{G}(X)\mathcal{I}_c(X,\theta_0)) & E_0(\mathbf{G}(X)\mathcal{I}_c(X,\theta_0)\mathbf{G}'(X)) \end{pmatrix} \end{aligned}$$

Now, elementary algebra gives

$$\Sigma_0 = \mathcal{I}(\theta_0)^{-1}.$$

Thus, using (18.38), we obtain

$$\begin{aligned}\sqrt{n}(\theta_n^* - \theta_0) \quad &\# \quad \mathcal{I}(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(y_i \mid x_i;\theta_0)}{\partial\theta} \\ &\# \quad \sqrt{n}(\hat{\theta}_n - \theta_0),\end{aligned}$$

where $\hat{\theta}_n$ denotes the maximum likelihood estimator of θ.

Moreover, because $\sqrt{n}(\hat{\theta}_n - \theta_n^*)$ converges to zero in probability, we can replace θ_n^* by $\hat{\theta}_n$ in the test statistic ξ_n^G. We obtain

$$\begin{aligned}\xi_n^G &= \frac{1}{n}\begin{pmatrix} 0 \\ \sum_{i=1}^n \mathbf{G}(x_i)\partial \log f(y_i \mid x_i;\hat{\theta}_n)/\partial\theta \end{pmatrix}' \\ &\quad \times (\hat{V}h)^{-1}\begin{pmatrix} 0 \\ \sum_{i=1}^n \mathbf{G}(x_i)\partial \log f(y_i \mid x_i;\hat{\theta}_n)/\partial\theta \end{pmatrix} \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{G}(x_i)\frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n)}{\partial\theta}\right)' \\ &\quad \times \hat{V}^{22}\left(\sum_{i=1}^{n}\mathbf{G}(x_i)\frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n)}{\partial\theta}\right),\end{aligned}$$

where $\hat{V}^{22}$ is a consistent estimator of the south-east block of $(Vh)^{-1}$. For instance, we can take

$$\begin{aligned}\hat{V}^{22} &= n\left(\sum_{i=1}^{n}\mathbf{G}(x_i)\hat{s}_i\hat{s}_i'\mathbf{G}'(x_i)\right. \\ &\quad \left. -\sum_{i=1}^{n}\mathbf{G}(x_i)\hat{s}_i\hat{s}_i'\left(\sum_{i=1}^{n}\hat{s}_i\hat{s}_i'\right)^{-1}\sum_{i=1}^{n}\hat{s}_i\hat{s}_i'\mathbf{G}'(x_i)\right)^{-1},\end{aligned}$$

where $\hat{s}_i = \partial \log f(y_i \mid x_i;\hat{\theta}_n)/\partial\theta$.

Hence we obtain

$$\begin{aligned}\xi_n^G = \sum_{i=1}^{n}\hat{s}_i'\mathbf{G}'(x_i)&\left(\sum_{i=1}^{n}\mathbf{G}(x_i)\hat{s}_i\hat{s}_i'\mathbf{G}'(x_i)\right. \\ &\left. -\sum_{i=1}^{n}\mathbf{G}(x_i)\hat{s}_i\hat{s}_i'\left(\sum_{i=1}^{n}\hat{s}_i\hat{s}_i'\right)^{-1}\sum_{i=1}^{n}\hat{s}_i\hat{s}_i'\mathbf{G}'(x_i)\right)^{-1}\sum_{i=1}^{n}\mathbf{G}'(x_i)\hat{s}_i.\end{aligned}$$

c) Linear Models

The preceding test statistic simplifies when the model is linear. For instance, consider the simple linear model

$$y_i = x_i\theta + u_i, \qquad i = 1, \ldots, n$$

where the y_i's and the x_i's are scalar and the u_i's are independently distributed $N(0,1)$ conditionally upon x_i.

We have

$$\frac{\partial \log f(y_i \mid x_i; \theta_n)}{\partial \theta} = x_i(y_i - x_i\theta).$$

Let $G(x_i) = x_i$. Then the test statistic becomes

$$\xi_n^G = \frac{\left(\sum_{i=1}^n \hat{u}_i x_i^2\right)^2 \sum_{i=1}^n \hat{u}_i^2 x_i^2}{\sum_{i=1}^n \hat{u}_i^2 x_i^4 \sum_{i=1}^n \hat{u}_i^2 x_i^2 - \left(\sum_{i=1}^n \hat{u}_i^2 x_i^3\right)^2},$$

where $\hat{u}_i = y_i - x_i\hat{\theta}_n$ and $\hat{\theta}_n$ is the OLS estimator of θ.

The critical region of the test is

$$\left\{\xi_n^G \geq \chi^2_{1-\alpha}(1)\right\}.$$

In this special case the test will detect a possible correlation between the error u_i and the square of the explanatory variable.

18.4 Hausman Specification Tests

18.4.1 Principle

Suppose that there exists a statistic τ_n such that

(i) under H_0, $\sqrt{n}\tau_n \xrightarrow{d} N(0, \Omega)$, where the rank d of Ω is constant under H_0,

(ii) under H_1, which is the complement of H_0, $\tau_n \to \tau_\infty$, where $\tau_\infty \neq 0$,

as $n \to \infty$.

Let $\hat{\Omega}_n^-$ be a consistent estimator of a generalized inverse Ω^- of Ω. Then, under the null hypothesis H_0, the test statistic $\xi_n^H = n\tau_n'\hat{\Omega}_n^-\tau_n$ satisfies $\xi_n^H \xrightarrow{d} \chi^2(d)$. Hence it is natural to consider the critical region

$$W = \left\{\xi_n^H \geq \chi^2_{1-\alpha}(d)\right\}.$$

An important special case arises when $\tau_n = T_n - T_n^0$, where

(i) T_n converges to $T(P)$ (say) for every probability distribution P of the model,

(ii) T_n^0 converges to $T(P)$ for every P under H_0, and T_n^0 is an asymptotically efficient estimator of $T(P)$ in the sense of having the minimal asymptotic variance covariance matrix in the class of estimators of the form

$$T_n^0 + \lambda(T_n - T_n^0), \quad \lambda \in I\!R,$$

(iii) $\sqrt{n}\left(T_n - T(P), T_n^0 - T(P)\right)$ is jointly asymptotically normally distributed under H_0,

as $n \to \infty$.

Under these conditions the asymptotic distribution of the random vector $\sqrt{n}\tau_n = \sqrt{n}(T_n - T_n^0)$ under the null hypothesis is the centered normal distribution with variance covariance matrix

$$V_{as}(T_n - T_n^0) = V_{as}T_n - V_{as}T_n^0. \tag{18.41}$$

The proof of such a result is identical to the proof of Property 6.10 dealing with unbiased estimation (see (6.1)).

Hence, assuming that the rank d of $V_{as}(T_n - T_n^0)$ remains constant under H_0, the critical region of the corresponding Hausman test is

$$W = \left\{ n(T_n - T_n^0)' \left(\hat{V}_{as}T_n - \hat{V}_{as}T_n^0\right)^{-} (T_n - T_n^0) \geq \chi^2_{1-\alpha}(d) \right\}. \tag{18.42}$$

18.4.2 Example: An Exogeneity Test Based on an Augmented Regression

a) The Model

We consider one equation of a simultaneous equations model, namely

$$Y_i = Y_{1i}'\beta_1 + X_{1i}'\beta_2 + u_i, \qquad i = 1, \ldots, n, \tag{18.43}$$

where Y_{1i} and X_{1i} are of dimension p_1 and k_1 respectively. It is assumed that the errors u_i are independent with zero mean and common variance σ^2. The vector X_{1i} is a subvector of the vector X_i of all exogenous variables of the system. The variables Y_{1i} are endogenous.

The null hypothesis of interest is that Y_{1i} is exogenous for the parameters in (18.43). This hypothesis can be written as

$$H_0 : \{E(Y_{1i}u_i) = 0,\ i = 1, \ldots, n\} \tag{18.44}$$

(see Exercise 18.5).

Stacking observations, equation (18.43) can be written in matrix form as

$$Y = \mathbf{Y}_1\beta_1 + \mathbf{X}_1\beta_2 + u, \tag{18.45}$$

or equivalently as

$$Y = \mathbf{Z}\beta + u \quad \text{with} \quad \mathbf{Z} = (\mathbf{Y}_1, \mathbf{X}_1). \tag{18.46}$$

b) Hausman Test Statistic

The ordinary least squares estimator of β is

$$\hat{\beta}_n^0 = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y. \tag{18.47}$$

The two-stage least squares estimator of β is

$$\hat{\beta}_n = (\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'Y, \tag{18.48}$$

where $\hat{\mathbf{Z}} = (\hat{\mathbf{Y}}_1, \mathbf{X}_1)$ is the matrix whose column vectors are the orthogonal projections of the column vectors of $\mathbf{Z}$ on the subspace generated by the columns of $\mathbf{X}$.

Under the null hypothesis of exogeneity, the Gauss Markov Theorem (or its asymptotic equivalent if there are some lagged endogenous variables in X_{1i}) implies that the OLS estimator $\hat{\beta}_n^0$ is efficient in the class of estimators of the form $\hat{\beta}_n^0 + \lambda(\hat{\beta}_n - \hat{\beta}_n^0)$. Hence the conditions for applying the Hausman test are satisfied. The Hausman test statistic is

$$\xi_n^H = \frac{1}{\hat{\sigma}_n^2}(\hat{\beta}_n - \hat{\beta}_n^0)' \left((\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1} - (\mathbf{Z}'\mathbf{Z})^{-1}\right)^{-} (\hat{\beta}_n - \hat{\beta}_n^0), \tag{18.49}$$

where $\hat{\sigma}_n^2$ is the estimator of σ^2 based on the OLS estimator of β.

c) Augmented Regression

We note that

$$\begin{aligned}
\hat{\mathbf{Z}}'\hat{\mathbf{Z}}(\hat{\beta}_n - \hat{\beta}_n^0) &= \hat{\mathbf{Z}}'Y - \hat{\mathbf{Z}}'\hat{\mathbf{Z}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y \\
&= \hat{\mathbf{Z}}'Y - \hat{\mathbf{Z}}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y \\
&= \hat{\mathbf{Z}}'(\mathbf{I} - \mathbf{P}_Z)Y,
\end{aligned}$$

where $\mathbf{P}_Z$ is the orthogonal projection matrix on the subspace generated by the columns of $\mathbf{Z} = (\mathbf{Y}_1, \mathbf{X}_1)$. Thus

$$(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})(\hat{\beta}_n - \hat{\beta}_n^0) = \begin{pmatrix} \hat{\mathbf{Y}}_1'(\mathbf{I} - \mathbf{P}_Z)Y \\ 0 \end{pmatrix}. \tag{18.50}$$

On the other hand, if we consider the auxiliary regression or so-called "augmented regression"

$$Y = \mathbf{Y}_1\beta_1 + \mathbf{X}_1\beta_2 + \hat{\mathbf{Y}}_1\gamma + \nu, \tag{18.51}$$

then ordinary least squares estimation of γ gives

$$\hat{\gamma}_n = \left(\hat{\mathbf{Y}}_1'(\mathbf{I} - \mathbf{P}_Z)\hat{\mathbf{Y}}_1\right)^{-1} \hat{\mathbf{Y}}_1'(\mathbf{I} - \mathbf{P}_Z)Y. \tag{18.52}$$

Comparing (18.50) and (18.52) shows that $\hat{\gamma}_n$ and $\hat{\beta}_n - \hat{\beta}_n^0$ are asymptotically in a bijective relationship through a determistic linear function. Hence the Hausman test based on $\hat{\beta}_n - \hat{\beta}_n^0$ is asymptotically equivalent to a test based on $\hat{\gamma}_n$.

The test statistic based on $\hat{\gamma}_n$ is

$$\xi_n = \frac{1}{\hat{\sigma}_n^2}\hat{\gamma}_n' \left(\hat{\mathbf{Y}}_1'(\mathbf{I} - \mathbf{P}_Z)\hat{\mathbf{Y}}_1\right)^{-1} \hat{\gamma}_n.$$

Its critical region is $\{\xi_n \geq \chi^2_{1-\alpha}(p_1)\}$.

Such a test is also asymptotically equivalent to the usual Fisher test of nullity of the coefficient γ in the augmented regression (18.51). Thus it can be easily performed using usual regression packages.

Note also that, if we replace $\hat{\mathbf{Y}}_1$ by $\mathbf{Y}_1 - \hat{\mathbf{Y}}_1$ in equation (18.51), then we obtain an estimator of γ that is equal to minus the preceding estimator of γ. Thus the same result obtains.

d) Tests Based on Estimators of β_1

From equation (18.50) we obtain immediately

$$(\hat{\beta}_{1n} - \hat{\beta}_{1n}^0) = (\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{11} \left(\hat{\mathbf{Y}}_1(\mathbf{I} - \mathbf{P}_Z)Y\right),$$

where $(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{11}$ is the block of the inverse of $\hat{\mathbf{Z}}'\hat{\mathbf{Z}}$ corresponding to the first p_1 rows and columns.

As a result, $\hat{\beta}_{1n} - \hat{\beta}^0_{1n}$ is in a bijective relationship with $\hat{\beta}_n - \hat{\beta}^0_n$. Hence we can consider only the subvector $\hat{\beta}_{1n} - \hat{\beta}^0_{1n}$ when constructing the test. Specifically the test statistic is

$$\frac{1}{\hat{\sigma}^2_n}(\hat{\beta}_{1n} - \hat{\beta}^0_{1n})' \left((\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{11} - (\mathbf{Z}'\mathbf{Z})^{11} \right)^{-1} (\hat{\beta}_{1n} - \hat{\beta}^0_{1n}).$$

18.5 Information Matrix Test

In this section we consider a given pseudo (or quasi) likelihood function, whose logarithm is

$$L_n(\theta) = \sum_{i=1}^{n} \log f(y_i \mid x_i; \theta),$$

where $\theta \in \Theta \subset \mathbb{R}^p$.

The null hypothesis of interest H_0 is that the true distribution generating the observations belongs to the family leading to the above likelihood function. Let $\theta_0 \in \Theta$ denote the true value of θ when H_0 holds.

The basic idea underlying the information matrix test relies on the equality between the two forms of the information matrix

$$\mathcal{I}(\theta_0) = E_x E_0 \left(\frac{\partial \log f(Y \mid X, \theta_0)}{\partial \theta} \, \frac{\partial \log f(Y \mid X, \theta_0)}{\partial \theta'} \right)$$

and

$$\mathcal{J}(\theta_0) = -E_x E_0 \left(\frac{\partial^2 \log f(Y \mid X; \theta_0)}{\partial \theta \partial \theta'} \right),$$

where E_0 denotes the expectation with respect to the true conditional distribution of Y given X when the null hypothesis holds, and on the fact that such an equality is no longer satisfied, in general, when H_0 does not hold.

Then the quantities of interest are the elements of the matrix $\mathcal{I}(\theta) - \mathcal{J}(\theta)$, where only the elements that are linearly independent are retained. Let q be the number of such quantities, which are indexed by ℓ so that $\ell = 1, \ldots, q$ with $q \leq p(p+1)/2$. The ℓ-th quantity corresponding to the (i, j)-th element is

$$a_\ell(y, x, \theta) = \frac{\partial \log f}{\partial \theta_i} \, \frac{\partial \log f}{\partial \theta_j} + \frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j}, \qquad \ell = 1, \ldots, q.$$

Let $a(y,x,\theta)$ denote the vector with components $a_\ell(y,x,\theta)$, $\ell = 1,\ldots,q$. Let

$$\hat{a}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} a(y_i,x_i,\hat{\theta}_n),$$

where $\hat{\theta}_n$ is the pseudo maximum likelihood estimator of θ.

Property 18.15: *Assume that the usual regularity conditions of the asymptotic theory of maximum likelihood estimation hold. Under H_0, the statistic $\sqrt{n}\hat{a}_n(\hat{\theta}_n)$ is asymptotically distributed as a centered normal distribution with variance covariance matrix*

$$\begin{aligned}\mathbf{V}(\theta_0) &= E_x E_0\left(a(Y,X,\theta_0) + \mathbf{A}(\theta_0)\mathcal{J}^{-1}(\theta_0)\frac{\partial \log f(Y \mid X;\theta_0)}{\partial\theta}\right)\\ &\quad\left(a(Y,X,\theta_0) + \mathbf{A}(\theta_0)\mathcal{J}^{-1}(\theta_0)\frac{\partial \log f(Y \mid X;\theta_0)}{\partial\theta}\right)',\end{aligned}$$

where

$$\mathbf{A}(\theta_0) = E_x E_0 \frac{\partial a(Y,X,\theta_0)}{\partial\theta'}.$$

PROOF: Taking a Taylor series expansion of $\sqrt{n}\hat{a}_n(\hat{\theta}_n)$ around θ_0, we obtain

$$\sqrt{n}\hat{a}_n(\hat{\theta}_n) \# \frac{1}{\sqrt{n}}\sum_{i=1}^{n} a(y_i,x_i,\theta_0) + \mathbf{A}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0).$$

Since

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \# \mathcal{J}^{-1}(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(y_i \mid x_i;\theta_0)}{\partial\theta},$$

we obtain

$$\sqrt{n}\hat{a}_n(\hat{\theta}_n) \# \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(a(y_i,x_i,\theta_0) + \mathbf{A}(\theta_0)\mathcal{J}^{-1}(\theta_0)\frac{\partial \log f(y_i \mid x_i;\theta_0)}{\partial\theta}\right). \tag{18.53}$$

The desired result follows from the Central Limit Theorem. □

The matrix $V(\theta_0)$ can be estimated consistently by

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}\left(a(y_i,x_i,\hat{\theta}_n)+\hat{\mathbf{A}}\hat{\mathcal{J}}^{-1}\frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n)}{\partial\theta}\right) \times\left(a(y_i,x_i,\hat{\theta}_n)+\hat{\mathbf{A}}\hat{\mathcal{J}}^{-1}\frac{\partial \log f(y_i \mid x_i;\hat{\theta}_n)}{\partial\theta}\right)', \quad (18.54)$$

where

$$\hat{\mathbf{A}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial a(y_i,x_i,\hat{\theta}_n)}{\partial\theta'}$$

and

$$\hat{\mathcal{J}} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(y_i \mid x_i;\hat{\theta}_n)}{\partial\theta\partial\theta'}.$$

Thus we obtain the following corollary.

Corollary 18.1: *Under H_0, if $\hat{\mathbf{V}}$ is nonsingular, then the test statistic*

$$\xi_n^I = n\hat{a}(\hat{\theta}_n)'\hat{V}^{-1}\hat{a}(\hat{\theta}_n)$$

is asymptotically distributed as a χ^2 distribution with q degrees of freedom.

PROOF: This is a direct consequence of the Property (18.15). □

Thus the test with critical region

$$W = \{\xi_n^I \geq \chi^2_{1-\alpha}(q)\} \quad (18.55)$$

is a test with asymptotic level α for testing the null hypothesis H_0. This test is called the *information matrix test.*

Example 18.1: Consider the linear model

$$y_i = \theta_1 + \theta_2 x_i + u_i, \qquad i = 1,\ldots,n,$$

where the u_i's are independently and normally distributed with zero mean and unit variance.

The conditional log-density for the i-th observation is

$$\log f(y_i \mid x_i;\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}(y_i-\theta_1-\theta_2 x_i)^2.$$

It follows that

$$\frac{\partial \log f(y_i \mid x_i;\theta)}{\partial \theta} = \begin{pmatrix} y_i - \theta_1 - \theta_2 x_i \\ x_i(y_i - \theta_1 - \theta_2 x_i) \end{pmatrix},$$

$$\frac{\partial^2 \log f(y_i \mid x_i;\theta)}{\partial \theta \partial \theta'} = \begin{pmatrix} -1 & -x_i \\ -x_i & -x_i^2 \end{pmatrix},$$

$$a(y_i \mid x_i, \theta) = \begin{pmatrix} (y_i - \theta_1 - \theta_2 x_i)^2 - 1 \\ x_i\left((y_i - \theta_1 - \theta_2 x_i)^2 - 1\right) \\ x_i^2\left((y_i - \theta_1 - \theta_2 x_i)^2 - 1\right) \end{pmatrix} = \begin{pmatrix} u_i^2 - 1 \\ x_i(u_i^2 - 1) \\ x_i^2(u_i^2 - 1) \end{pmatrix},$$

and

$$\hat{a}(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \hat{u}_i^2 - 1 \\ x_i(\hat{u}_i^2 - 1) \\ x_i^2(\hat{u}_i^2 - 1) \end{pmatrix},$$

where $\hat{u}_i = y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i$, and $\hat{\theta}_1$ and $\hat{\theta}_2$ are the ordinary least squares estimators of θ_1 and θ_2. Hence the information matrix test reduces to testing whether the variance of u_i is equal to 1 and whether u_i^2 is uncorrelated with x_i and x_i^2.

The test statistic is easily computed from the expressions for a, $\partial \log f/\partial\theta$, and $\partial^2 \log f/\partial\theta\partial\theta'$ given above, and the expression for $\hat{\mathbf{A}}$, which is

$$\hat{\mathbf{A}} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \\ x_i^2 & x_i^3 \end{pmatrix}.$$

From the normal equations of OLS estimation, we also have

$$\hat{\mathbf{A}} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i) \begin{pmatrix} 0 & 0 \\ 0 & x_i^2 \\ x_i^2 & x_i^3 \end{pmatrix}.$$

Example 18.2: Consider the model

$$y_i = \theta_1 + u_i,$$

where the u_i's are independently and normally distributed with zero mean and common variance θ_2. The log-likelihood for the i-th observation is

$$\log f(y_i;\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \theta_2 - \frac{1}{2\theta_2}(y_i - \theta_1)^2.$$

Hence

$$\frac{\partial \log f(y_i;\theta)}{\partial \theta} = \frac{1}{\theta_2}\begin{pmatrix} y_i-\theta_1 \\ -\frac{1}{2}+\frac{1}{2\theta_2}(y_i-\theta_1)^2 \end{pmatrix},$$

$$\frac{\partial^2 \log f(y_i;\theta)}{\partial\theta\partial\theta'} = -\frac{1}{\theta_2}\begin{pmatrix} 1 & \frac{y_i-\theta_1}{\theta_2} \\ \frac{y_i-\theta_1}{\theta_2} & -\frac{1}{2\theta_2}+\frac{1}{\theta_2^2}(y_i-\theta_1)^2 \end{pmatrix},$$

$$a(y_i,x_i,\theta) = \frac{1}{\theta_2}\begin{pmatrix} \frac{1}{\theta_2}(y_i-\theta_1)^2-1 \\ -\frac{1}{2}(y_i-\theta_1)+\frac{1}{2\theta_2}(y_i-\theta_1)^3-\frac{1}{\theta_2}(y_i-\theta_1) \\ \frac{1}{4\theta_2}+\frac{1}{4\theta_2^2}(y_i-\theta_1)^4-\frac{1}{2\theta_2^2}(y_i-\theta_1)^2+\frac{1}{2\theta_2}-\frac{1}{\theta_2^2}(y_i-\theta_1)^2 \end{pmatrix},$$

and

$$\hat{a}_n(\hat{\theta}_n) = \frac{1}{\hat{\theta}_2}\begin{pmatrix} 0 \\ \frac{1}{2\hat{\theta}_2}\sum_{i=1}^n (y_i-\hat{\theta}_1)^3 \\ -\frac{3}{4\hat{\theta}_2}+\frac{1}{4\hat{\theta}_2^2}\sum_{i=1}^n (y_i-\hat{\theta}_1)^4 \end{pmatrix}.$$

Thus the information matrix test reduces to testing whether the skewness and kurtosis coefficients $E(u_i^3)/\theta_2^{3/2}$ and $E(u_i^4)/\theta_2^2 - 3$ are equal to zero as it is the case for the normal distribution.

18.6 Exercises

EXERCISE 18.1: Consider the mixed hypothesis about the parameter θ defined by

$$H_0 : \{\exists a \in I\!R^q : Aa + B\theta + C = 0\},$$

where A, B, and C are given constants. Interpret the rank condition (18.6).

EXERCISE 18.2 : Verify that the minimization problem

$$\min_a \Phi(a) = n\, g'(\hat{\theta}_n, a)\ \mathbf{S}_n\ g(\hat{\theta}_n, a) \tag{18.23}$$

can also be written as

$$\begin{cases} \min_{\theta,a} n g'(\theta,a)\mathbf{S}_n g(\theta,a) \\ \text{subject to } \theta = \hat{\theta}_n. \end{cases}$$

Define the test statistic

$$\xi_n^L = \frac{1}{n}\hat{\mu}_n \hat{V}_{as}\left(\frac{\hat{\mu}_n}{\sqrt{n}}\right)^- \hat{\mu}_n,$$

where $\hat{\mu}_n$ is the Lagrange multiplier associated with the constraint $\theta = \hat{\theta}_n$. Conclude that ξ_n^L is asymptotically equivalent to the test statistics considered in Section 18.1

EXERCISE 18.3: Consider the model (18.31), i.e.

$$y_t = \beta z_t + \theta_1 x_t + \theta_2 x_{t-1} + \cdots + \theta_p x_{t-p+1} + u_t, \qquad t = 1, \ldots, T,$$

where $\Theta(L) = \theta_1 + \theta_2 L + \ldots + \theta_p L^{p-1}$ is the expansion of the rational fraction $C(L)/D(L)$ truncated at the order $p-1$. Let the degrees of the polynomials C and D be m and n respectively, with $D(0) = 1$.

a) Show that the θ_i's satisfy the condition

$$\exists d_1, \ldots, d_n : \theta_i + d_1\theta_{i-1} + \ldots + d_n\theta_{i-n} = 0, \ \forall i \geq m+1,$$

with the convention $\theta_i = 0$ for $i < 0$.

b) Propose a general method for testing the hypothesis that lags are rational. Hint: Generalize the approach of Section 18.2.3-a).

EXERCISE 18.4: Consider the simultaneous equations model

$$\mathbf{B}y_t + \mathbf{C}x_t = u_t, \qquad t = 1, \ldots, T,$$

where y_t is a vector of endogenous variables, x_t is a vector of exogenous variables, and $\mathbf{B}$ and $\mathbf{C}$ are matrices subject to some exclusion restrictions. Let $y_t = \Pi x_t + v_t$ denote the unconstrained reduced form.

a) Suppose that the model is overidentified. Write the overidentifying constraints under a mixed form.

b) Propose a test of the overidentifying restrictions.

EXERCISE 18.5: Consider equation (18.43), i.e.

$$Y_i = Y_{1i}'\beta_1 + X_{1i}'\beta_2 + u_i, \qquad i = 1, \ldots, n,$$

which is one equation of a simultaneous equations model. The reduced form equations for the vector Y_{1i} are

$$Y_{1i} = \Pi X_i + v_i,$$

where X_i is the vector of all exogenous variables.

Suppose that the vectors (u_i, v_i) are independently and identically normally distributed with zero mean. Show that Y_{1i} is exogenous for the parameters of equation (18.43) in the sense of Definition 1.8 if hypothesis (18.44) holds, i.e., if $E(Y_{1i}u_i) = 0$.

EXERCISE 18.6 : Consider the dynamic model

$$\begin{cases} y_t &= dy_{t-1} + x_t\beta + \nu_t, \\ \nu_t &= \rho\nu_{t-1} + \epsilon_t, \end{cases}$$

for $t = 1, \ldots, T$, where x_t is a row vector of exogenous variables, $\mid \rho \mid < 1$ and the ϵ_t's are independent with zero mean and variance σ^2.

Show that the model can be written as

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + x_t\theta_3 + x_{t-1}\theta_4 + \epsilon_t,$$

where the parameters θ_i's are constrained by a condition of the form (18.30). Then propose a procedure for testing such a constraint.

EXERCISE 18.7: Consider the model

$$y_t = \theta_1 y_{t-1} + \theta_2 x_t + \theta_3 x_{t-1} + \epsilon_t,$$

where the ϵ_t's are independent with zero mean and variance σ^2. How would you test the hypothesis that the model is

$$y_t = ax_t + u_t,$$

where u_t follows an autoregressive process of order 1?

EXERCISE 18.8: Consider the linear model

$$y_i = \theta_1 + \theta_2 x_i + u_i,$$

where the u_i's are independently and normally distributed with zero mean and variance θ_3. Derive and interpret the information matrix test.

18.7 References

Burguete, J.F., Gallant, A.R., and Souza, G. (1982). "On the Unification of the Asymptotic Theory of Non Linear Econometric Models," *Econometric Review* , 1, 151–190.

Gallant, A.R. (1987). *Non Linear Statistical Models*, Wiley.

Gallant, A.R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Non Linear Dynamic Models*, Basil Blackwell.

Gourieroux, C. and Monfort, A. (1989). "A General Framework for Testing a Null Hypotheses in a Mixed Form ," *Econometric Theory*, 5, 63–82.

Hausman, J. (1978). "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1272.

Le Cam, L.(1986). *Asymptotic Methods in Statistical Decision*, Springer Verlag.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, Wiley.

Szroeter, J. (1983). "Generalized Wald Methods for Testing Non Linear Implicit and Overidentifying Restrictions," *Econometrica*, 51, 335–348.

White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

White, H. (1934). *Asymptotic Theory for Econometricians*, Academic Press.

CHAPTER 19

Multiple Tests

In this chapter we are interested in more than one testing problem. First, we shall consider the case of a decreasing nested sequence of null hypotheses (in implicit or explicit forms). Then, we shall examine the case where the null hypotheses in this sequence are nonnested. Lastly, we shall study the case of testing a null hypothesis that is nested in more than one model.

To simplify the exposition, we shall consider the fully parametric case where a likelihood function is available and where the null hypotheses are defined in either implicit or explicit forms. The results presented in this chapter can be extended to the general framework of Chapter 18, i.e., to multiple testing of hypotheses in mixed forms in semiparametric models. It is, however, important to mention that the asymptotic independence that is established in this chapter remains valid within this more general framework, provided the additional hypothesis $\mathcal{I}(\theta) = \mathcal{J}(\theta)$ is satisfied (see Exercises 19.1 and 19.2). Note that such a hypothesis was already made in the preceding chapter for the generalization of the likelihood-based tests of Chapter 17.

We begin with the linear model.

19.1 The Linear Model

19.1.1 Nested Hypotheses

We consider the linear model defined by

$$Y = X_1 b_1 + \ldots + X_K b_K + u, \tag{19.1}$$

where Y, $X_1, \ldots, X_K$ and u are vectors of size n. The coefficients $b_1, \ldots, b_K$ are unknown. It is assumed that the X_K's, $k = 1, \ldots, K$, are nonstochastic and that the matrix $\mathbf{X} = (X_1, \ldots, X_K)$ is of rank equal

to K. It is also assumed that the distribution of u is the multivariate normal distribution $N(0, \sigma^2\mathbf{I})$, where σ^2 is unknown.

We consider the following hypotheses:

$$\begin{array}{lll} H_0^{K+1} & : & \text{the unconstrained model (by convention)}, \\ H_0^K & : & b_K = 0, \\ H_0^{K-1} & : & b_K = b_{K-1} = 0, \\ \vdots & & \\ H_0^k & : & b_K = b_{K-1} = \ldots = b_k = 0, \\ \vdots & & \\ H_0^1 & : & b_K = b_{K-1} = \ldots = b_1 = 0. \end{array}$$

These hypotheses define a decreasing sequence of subsets of the parameter space. Thus each hypothesis, if true, implies the validity of all preceding ones.

19.1.2 The Testing Procedure

We wish to test these hypotheses successively. To do so, we start by testing the null hypothesis H_0^K against the unconstrained model H_0^{K+1}. If H_0^K is rejected, the testing procedure ends. Otherwise, the null hypothesis H_0^{K-1} is tested against the model constrained by H_0^K. The testing procedure continues until a null hypothesis H_0^k is rejected, in which case H_0^{k+1} is the smallest null hypothesis that is accepted, or until H_0^1 is accepted. If W_k denotes the critical region of the test for testing H_0^k against the model constrained by H_0^{k+1}, then the sequential testing procedure thus defined consists in accepting H_0^k when the observation y belongs to $\bigcap_{j=k}^K {}^cW_j$.

Now, for each k, the test of the null hypothesis H_0^k against the model constrained by H_0^{k+1} reduces to a test of the null hypothesis $b_k = 0$ in the model

$$Y = X_1 b_1 + \ldots + X_k b_k + u. \tag{19.2}$$

Then, from Chapter 16 we know that for such a testing problem and any given level α_k, there exists a test that is UMPU and UMPI. This test is the two-sided Student test, or equivalently, the Fisher test, whose critical region is

$$W_k = \left\{ F_k = \frac{\|\mathbf{M}_k Y - \mathbf{M}_{k-1} Y\|^2}{\|(\mathbf{I} - \mathbf{M}_k)Y\|^2/(n-k)} \geq F_{1-\alpha_k}(1, n-k) \right\},$$

where $\mathbf{M}_k$ denotes the orthogonal projection of $\mathbb{R}^n$ on to the subspace generated by the vectors $X_1, \ldots, X_k$.

The critical regions W_k, or equivalently the statistics F_k, have an interesting independence property, which is formally stated in the following property.

Property 19.1: *Under H_0^k, the statistics $F_K, \ldots, F_k$ and $(F_{k-1}, \ldots, F_1)$ are mutually independent.*

PROOF:

a) Under H_0^k, the statistic F_k is ancillary since its probability distribution, which is the Fisher–Snedecor distribution with one and $n-k$ degrees of freedom, does not depend on the parameters $b_1, \ldots, b_{k-1}, \sigma^2$.

b) On the other hand, it can be shown (see Exercise 19.3) that under H_0^k, the statistics $(\mathbf{M}_{k-1}Y, \|(\mathbf{I}-\mathbf{M}_{k-1})Y\|^2)$ is complete and sufficient. Thus, from Property 3.29, the statistic F_k is independent of $(\mathbf{M}_{k-1}Y, \|(\mathbf{I}-\mathbf{M}_{k-1})Y\|^2)$.

Moreover, the statistics F_{k-i}, $i \geq 1$ can be written as

$$\begin{aligned} F_{k-i} &= \frac{\|\mathbf{M}_{k-i}Y - \mathbf{M}_{k-i-1}Y\|^2}{\|(\mathbf{I}-\mathbf{M}_{k-i})Y\|^2/(n-k+i)} \\ &= \frac{\|\mathbf{M}_{k-i}\mathbf{M}_{k-1}Y - \mathbf{M}_{k-i-1}\mathbf{M}_{k-1}Y\|^2}{\big(\|(\mathbf{I}-\mathbf{M}_{k-1})Y\|^2 + \|\mathbf{M}_{k-1}Y - \mathbf{M}_{k-i}\mathbf{M}_{k-1}Y\|^2\big)/(n-k+i)}. \end{aligned}$$

For every $i \geq 1$, it follows that F_{k-i} is a function of the pair of statistics $(\mathbf{M}_{k-1}Y, \|(\mathbf{I}-\mathbf{M}_{k-1})Y\|^2)$. Hence, under H_0^k, the statistic F_k is independent of $(F_{k-1}, \ldots, F_1)$.

c) Now note that H_0^k implies $H_0^K, \ldots, H_0^{k+1}$. Therefore, by applying b) to each of these hypotheses, we have under H_0^k:

$$\begin{array}{lll} F_K & \text{independent of} & (F_{K-1}, \ldots, F_1), \\ F_{K-1} & \text{independent of} & (F_{K-2}, \ldots, F_1), \\ & \vdots & \\ F_k & \text{independent of} & (F_{k-1}, \ldots, F_1). \end{array}$$

Or equivalently, the statistics $F_K, F_{K-1}, \ldots, F_k$ and $(F_{k-1}, \ldots, F_1)$ are mutually independent. □

Remark 19.1: Property 19.1 implies the mutual independence of the statistics $F_K, \ldots, F_k$, and F_{k-i}, for every $i \geq 1$ under H_0^k. It does not,

however, imply the mutual independence of the statistics $F_K, \ldots, F_k$, F_{k-i}, F_{k-j}, for $i \neq j$, $i \geq 1$, $j \geq 1$, or even of F_{k-i}, and F_{k-j}, for $i \geq 1$, $j \geq 1$.

Property 19.1 can be used in particular to compute the risk of Type I error, i.e., the significance level for testing H_0^k as a function of the chosen levels $\alpha_K, \ldots, \alpha_{k+1}$ of the preceding individual tests. Specifically

$$\begin{aligned} P(\text{reject } H_0^k \mid H_0^k \text{ is true}) &= 1 - P(\text{accept } H_0^k \mid H_0^k \text{ is true}) \\ &= 1 - P\left(\bigcap_{j=k}^{K} {}^cW_j \mid H_0^k \text{ is true}\right). \end{aligned}$$

Thus

$$\begin{aligned} &P(\text{reject } H_0^k \mid H_0^k \text{ is true}) = 1- \\ &\quad P(F_K < F_{1-\alpha_K}(1, n-K), \ldots, F_k < F_{1-\alpha_k}(1, n-k) \mid H_0^k \text{ is true}). \end{aligned}$$

Since the statistics $F_K, \ldots, F_k$ are independent under H_0^k, we obtain

$$P(\text{reject } H_0^k \mid H_0^k \text{ is true}) = 1 - \prod_{j=k}^{K}(1-\alpha_j), \tag{19.3}$$

for every $k = 1, \ldots, K$.

Thus, an appropriate choice for $\alpha_j, j = 1, \ldots, K$, can give the a priori chosen levels of significance for testing the null hypotheses H_0^k, $k = 1, \ldots, K$, provided the latter desired levels increase as k decreases. Conversely, a priori chosen levels $\alpha_j, j = 1, \ldots, K$, induce significance levels for testing the null hypotheses H_0^k that are increasing as k decreases. These significance levels are given by (19.3). In particular, if all the α_j's are equal to α, the induced significance level for testing H_0^k is

$$1 - (1-\alpha)^{K-k+1}, \tag{19.4}$$

for every $k = 1, \ldots, K$.

The fact that individual significance levels can be controlled is an interesting property of the sequential testing procedure defined above. Such a procedure is sometimes qualified as a *descending stepwise procedure*. This procedure is in opposition to another procedure, which is an *ascending stepwise procedure*. The latter procedure consists of starting

from a relatively small model including (say) the first k_0 explanatory variables (with possibly $k_0 = 0$) and of testing successively the inclusion of an additional explanatory variable. The ascending stepwise procedure ends when an additional variable is not accepted. Unlike the descending stepwise procedure, a drawback of the ascending stepwise procedure is that the significance level for testing H_0^k cannot be controlled. For instance, assuming that $k_0 = 0$, this procedure gives

$$\begin{aligned} & P(\text{reject } H_0^k \mid H_0^k \text{ is true }) \\ &= P\left(\bigcap_{j=1}^{k} W_j \mid H_0^k \text{ is true }\right) \\ &= P(F_1 \geq F_{1-\alpha_1}(1, n-1), \ldots, F_k \geq F_{1-\alpha_k}(1, n-k) \mid H_0^k \text{ is true }). \end{aligned}$$

Now, the statistics that are involved when H_0^k is rejected are $F_1, \ldots, F_k$. Though the proof of Property 19.1 shows that the statistic F_k is independent of $(F_{k-1}, \ldots, F_1)$ for $k \geq 1$ under H_0^k, the statistics F_i, $1 \geq i \geq k-1$ are not mutually independent.

19.1.3 Interpretation in Terms of Correlation Coefficients

The preceding testing procedure is related to a more descriptive procedure based on the *empirical partial correlation coefficient* R_k between Y and X_k given $X_1, \ldots, X_{k-1}$. This follows from the fact that the test statistic F_k is an increasing function of R_k.

Specifically

$$\begin{aligned} \frac{1}{F_k} &= \frac{1}{n-k}\left(\frac{\|(\mathbf{I}-\mathbf{M}_k)Y\|^2}{\|\mathbf{M}_k Y - \mathbf{M}_{k-1}Y\|^2}\right) \\ &= \frac{1}{n-k}\frac{\|(\mathbf{I}-\mathbf{M}_{k-1})Y\|^2 - \|\mathbf{M}_k Y - \mathbf{M}_{k-1}Y\|^2}{\|\mathbf{M}_k Y - \mathbf{M}_{k-1}Y\|^2} \\ &= \frac{1}{n-k}\left(\frac{\|(\mathbf{I}-\mathbf{M}_{k-1})Y\|^2}{\|\mathbf{M}_k Y - \mathbf{M}_{k-1}Y\|^2} - 1\right). \end{aligned}$$

But

$$\|\mathbf{M}_k Y - \mathbf{M}_{k-1}Y\|^2 = \frac{(X_k'(\mathbf{I}-\mathbf{M}_{k-1})Y)^2}{X_k'(\mathbf{I}-\mathbf{M}_{k-1})X_k}.$$

Hence

$$\begin{aligned}\frac{1}{F_k} &= \frac{1}{n-k}\left(\frac{Y'(\mathrm{I}-\mathrm{M}_{k-1})YX_k'(\mathrm{I}-\mathrm{M}_{k-1})X_k}{(X_k'(\mathrm{I}-\mathrm{M}_{k-1})Y)^2}-1\right)\\ &= \frac{1}{n-k}\left(\frac{1}{R_k^2}-1\right).\end{aligned}$$

Thus we have the relation

$$F_k = (n-k)\frac{R_k^2}{1-R_k^2}. \tag{19.5}$$

19.2 Sequences of Null Hypotheses in Implicit Form: The General Case

We assume that the observed variables can be partitioned into endogenous variables denoted Y and exogenous variables denoted X. The observations are (Y_i, X_i), $i = 1, \ldots, n$. To simplify, we also assume that the pairs (Y_i, X_i) are independently and identically distributed.

The parameter vector of interest θ belongs to an open subset Θ of $\mathbb{R}^p$. It parameterizes the conditional distribution of Y_i given X_i. The density is $f(y_i \mid x_i; \theta)$. The conditional log-likelihood function is denoted $L_n(\theta) = \sum_{i=1}^n \log f(y_i \mid x_i; \theta)$.

We first generalize the results of Chapter 17 on Wald, score, and likelihood ratio tests to the case where the general or maintained hypothesis already imposes some constraints on the parameter θ. Then we apply these results to the problem of testing a sequence of null hypotheses in implicit form.

19.2.1 Test of a Hypothesis When the Maintained Hypothesis is Already Constrained

We consider the case where the parameter vector θ satisfies the constraint $g^*(\theta) = 0$ under the maintained hypothesis, where g^* is a r^*-dimensional function. We want to test the null hypothesis H_0 that θ satisfies the constraint $g^{**}(\theta) = 0$, where g^{**} is a r^{**}-dimensional function. Let

$$g(\theta) = \begin{pmatrix} g^*(\theta) \\ g^{**}(\theta) \end{pmatrix},$$

where $g(\theta)$ is of size $r = r^* + r^{**}$. It is assumed that the Jacobian matrix $\partial g'/\partial\theta$ is of rank r.

Let $\hat{\theta}_n^*$ be the maximum likelihood estimator of θ under the maintained hypothesis $g^*(\theta) = 0$. Let $\hat{\theta}_n^0$ denote the maximum likelihood estimator of θ under the null hypothesis $H_0 : g(\theta) = 0$.

To test the null hypothesis H_0, we shall generalize the results of Chapter 17 concerning Wald, score (or Lagrange multiplier), and likelihood ratio tests. Specifically, since θ must satisfy some constraints both under the null and maintained hypotheses, we need to distinguish two sets of Lagrange multipliers. We let $\hat{\mu}_n$ denote the vector of Lagrange multipliers corresponding to the maximization of the log-likelihood under the constraints of the maintained hypothesis. Similarly, we let $(\hat{\mu}_n^{0\prime}, \hat{\lambda}_n')'$ denote the vector of Lagrange multipliers corresponding to the maximization of the log-likelihood under the constraints of the null hypothesis H_0. These multipliers must satisfy the following first-order conditions:

$$\frac{\partial L_n(\hat{\theta}_n^*)}{\partial\theta} + \frac{\partial g^{*\prime}(\hat{\theta}_n^*)}{\partial\theta}\hat{\mu}_n = 0, \tag{19.6}$$

$$\frac{\partial L_n(\hat{\theta}_n^0)}{\partial\theta} + \frac{\partial g^{*\prime}(\hat{\theta}_n^0)}{\partial\theta}\hat{\mu}_n^0 + \frac{\partial g^{**\prime}(\hat{\theta}_n^0)}{\partial\theta}\hat{\lambda}_n = 0. \tag{19.7}$$

Property 19.2: *Within the preceding framework we have*

$$\xi_n^W \# \xi_n^S \# \xi_n^R$$

under H_0, *where*

$$\begin{aligned}
\xi_n^W &= n g^{**\prime}(\hat{\theta}_n^*)\left(\frac{\partial g^{**}(\hat{\theta}_n^*)}{\partial\theta'}\mathbf{M}_*'(\hat{\theta}_n^*)\mathcal{I}^{-1}(\hat{\theta}_n^*)\mathbf{M}_*(\hat{\theta}_n^*)\frac{\partial g^{**\prime}(\hat{\theta}_n^*)}{\partial\theta}\right)^{-1} \\
&\quad \times g^{**}(\hat{\theta}_n^*), \\
\xi_n^S &= \frac{1}{n}\hat{\lambda}_n'\frac{\partial g^{**}(\hat{\theta}_n^0)}{\partial\theta'}\mathbf{M}_*'(\hat{\theta}_n^0)\mathcal{I}^{-1}(\hat{\theta}_n^0)\mathbf{M}_*(\hat{\theta}_n^0)\frac{\partial g^{**\prime}(\hat{\theta}_n^0)}{\partial\theta}\hat{\lambda}_n \\
&= \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial\theta'}\mathbf{M}_*'(\hat{\theta}_n^0)\mathcal{I}^{-1}(\hat{\theta}_n^0)\mathbf{M}_*(\hat{\theta}_n^0)\frac{\partial L_n(\hat{\theta}_n^0)}{\partial\theta}, \\
\xi_n^R &= 2\left(L_n(\hat{\theta}_n^*) - L_n(\hat{\theta}_n^0)\right),
\end{aligned}$$

and

$$\mathbf{M}_*(\theta) = \mathbf{I} - \frac{\partial g^{*\prime}(\theta)}{\partial \theta}\left(\frac{\partial g^*(\theta)}{\partial \theta'}\mathcal{I}^{-1}(\theta)\frac{\partial g^{*\prime}(\theta)}{\partial \theta}\right)^{-1}\frac{\partial g^*(\theta)}{\partial \theta'}\mathcal{I}^{-1}(\theta).$$

Moreover, the three test statistics ξ_n^W, ξ_n^S, *and* ξ_n^R *are asymptotically* $\chi^2(r^{**})$ *distributed under* H_0. *Thus the tests with critical regions* $\{\xi_n \geq \chi^2_{1-\alpha}(r^{**})\}$, *where* ξ_n *is any one of the above three test statistics, are asymptotically of level* α *and are consistent.*

PROOF:

a) ASYMPTOTIC RELATIONS BETWEEN CONSTRAINED AND UNCONSTRAINED ESTIMATORS:

Let θ_0 denote the true parameter value of θ. Let $\hat{\theta}_n$ be the unconstrained maximum likelihood estimator of θ. We have

$$0 = \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{19.8}$$

Moreover, we have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^*)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^* - \theta_0). \tag{19.9}$$

Taking the difference, we obtain

$$\mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^* - \theta_0) \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) - \frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^*)}{\partial \theta}. \tag{19.10}$$

Then, using (19.6), we obtain

$$\mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^* - \theta_0) \# \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + \frac{\partial g^{*\prime}(\hat{\theta}_n^*)}{\partial \theta}\frac{\hat{\mu}_n}{\sqrt{n}}. \tag{19.11}$$

On the other hand, from (17.12) we have

$$\frac{\hat{\mu}_n}{\sqrt{n}} \# -\left(\frac{\partial g^*(\theta_0)}{\partial \theta'}\mathcal{I}^{-1}(\theta_0)\frac{\partial g^{*\prime}(\theta_0)}{\partial \theta}\right)^{-1}\sqrt{n}g^*(\hat{\theta}_n).$$

Hence

$$\frac{\hat{\mu}_n}{\sqrt{n}} \# -\left(\frac{\partial g^*(\theta_0)}{\partial \theta'}\mathcal{I}^{-1}(\theta_0)\frac{\partial g^{*\prime}(\theta_0)}{\partial \theta}\right)^{-1}\sqrt{n}\frac{\partial g^*(\theta_0)}{\partial \theta'}(\hat{\theta}_n - \theta_0). \tag{19.12}$$

Hence (19.11) can be written as

$$\sqrt{n}(\hat{\theta}_n^* - \theta_0) \# \left(\mathbf{I} - \mathcal{I}^{-1}(\theta_0) \frac{\partial g^{*\prime}(\theta_0)}{\partial \theta} \left(\frac{\partial g^*(\theta_0)'}{\partial \theta} \mathcal{I}^{-1}(\theta_0) \frac{\partial g^*(\theta_0)'}{\partial \theta} \right)^{-1} \frac{\partial g^*(\theta_0)}{\partial \theta'} \right) \times \sqrt{n}(\hat{\theta}_n - \theta_0), \tag{19.13}$$

or equivalently as

$$\sqrt{n}(\hat{\theta}_n^* - \theta_0) \# \mathbf{M}_*'(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0), \tag{19.14}$$

where $\mathbf{M}_*(\theta_0)$ is the matrix of the orthogonal projection matrix on to the subspace orthogonal to the subspace generated by the columns of $\partial g^{*\prime}(\theta_0)/\partial \theta$ with respect to the scalar product defined by $\mathcal{I}^{-1}(\theta_0)$.

b) ASYMPTOTIC DISTRIBUTION OF ξ_n^W:
From (19.14) it follows that, under H_0,

$$\sqrt{n} g^{**}(\hat{\theta}_n^*) \# \frac{\partial g^{**}(\theta_0)}{\partial \theta'} \mathbf{M}_*'(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0). \tag{19.15}$$

Hence

$$\sqrt{n} g^{**}(\hat{\theta}_n^*) \xrightarrow{d} N\left(0, \frac{\partial g^{**}(\theta_0)}{\partial \theta'} \mathbf{M}_*'(\theta_0) \mathcal{I}^{-1}(\theta_0) \mathbf{M}_*(\theta_0) \frac{\partial g^{**'}(\theta_0)}{\partial \theta}\right). \tag{19.16}$$

It follows that the asymptotic distribution of the Wald test statistic ξ_n^W is a $\chi^2(r^{**})$ distribution under H_0. The corresponding Wald test follows.

c) ASYMPTOTIC RELATIONS AMONG LAGRANGE MULTIPLIERS:
From the definitions of the constrained estimators $\hat{\theta}_n^*$ and $\hat{\theta}_n^0$ we have

$$\begin{aligned} 0 &= \sqrt{n} g^*(\hat{\theta}_n^*) = \sqrt{n} g^*(\hat{\theta}_n^0) \\ &\# \frac{\partial g^*(\theta_0)}{\partial \theta'} \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \\ &\# \frac{\partial g^*(\theta_0)}{\partial \theta'} \mathcal{I}^{-1}(\theta_0) \frac{1}{\sqrt{n}} \left(\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} - \frac{\partial L_n(\hat{\theta}_n^*)}{\partial \theta} \right). \end{aligned}$$

Then, using (19.6) and (19.7), we obtain

$$0 \# \frac{\partial g^*(\theta_0)}{\partial \theta'} \mathcal{I}^{-1}(\theta_0) \frac{1}{\sqrt{n}} \left(\frac{\partial g^{*\prime}(\theta_0)}{\partial \theta} (\hat{\mu}_n - \hat{\mu}_n^0) - \frac{\partial g^{**\prime}(\theta_0)}{\partial \theta} \hat{\lambda}_n \right).$$

Thus we have the following asymptotic relation between $\hat{\lambda}_n/\sqrt{n}$ and $(\hat{\mu}_n - \hat{\mu}_n^0)/\sqrt{n}$:

$$\frac{\hat{\mu}_n - \hat{\mu}_n^0}{\sqrt{n}} \;\#\; \left(\frac{\partial g^*(\theta_0)}{\partial \theta'}\mathcal{I}^{-1}(\theta_0)\frac{\partial g^{*\prime}(\theta_0)}{\partial \theta}\right)^{-1} \cdot \times \frac{\partial g^*(\theta_0)}{\partial \theta'}\mathcal{I}^{-1}(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}}. \tag{19.17}$$

d) Asymptotic Equivalence Between ξ_n^W and ξ_n^S:
We have

$$\sqrt{n}g^{**}(\hat{\theta}_n^*) \;\#\; \frac{\partial g^{**}(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^* - \theta_0)$$

and

$$0 = \sqrt{n}g^{**}(\hat{\theta}_n^0) \;\#\; \frac{\partial g^{**}(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0).$$

Taking the difference, we obtain

$$\sqrt{n}g^{**}(\hat{\theta}_n^*) \;\#\; \frac{\partial g^{**}(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0). \tag{19.18}$$

But

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \;\#\; \frac{1}{\sqrt{n}}\mathcal{I}^{-1}(\theta_0)\left(\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} - \frac{\partial L_n(\hat{\theta}_n^*)}{\partial \theta}\right),$$

which gives using (19.6) and (19.7)

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \;\#\; \frac{1}{\sqrt{n}}\mathcal{I}^{-1}(\theta_0)\frac{\partial g'(\theta_0)}{\partial \theta}\begin{pmatrix} \hat{\mu}_n - \hat{\mu}_n^0 \\ -\hat{\lambda}_n \end{pmatrix}.$$

Hence, using (19.17) and omitting θ_0, we have

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \;\# \\ -\mathcal{I}^{-1}(\theta_0)\left(\mathbf{I} - \frac{\partial g^{*\prime}}{\partial \theta}\left(\frac{\partial g^*}{\partial \theta'}\mathcal{I}^{-1}\frac{\partial g^{*\prime}}{\partial \theta}\right)^{-1}\frac{\partial g^*}{\partial \theta'}\mathcal{I}^{-1}\right)\frac{\partial g^{**\prime}}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}},$$

i.e.

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \;\#\; -\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}}. \tag{19.19}$$

Thus, using (19.18) and (19.19), we obtain

$$\sqrt{n}g^{**}(\hat{\theta}_n^*) \# -\frac{\partial g^{**}(\theta_0)}{\partial \theta'}\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}}.$$

Now, because $\mathbf{M}_*(\theta_0)$ is the matrix of an orthogonal projection, we have the equality

$$\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0) = \mathbf{M}_*'(\theta_0)\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0).$$

Hence

$$\sqrt{n}g^{**}(\hat{\theta}_n^*) \# -\frac{\partial g^{**}(\theta_0)}{\partial \theta'}\mathbf{M}_*'(\theta_0)\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}}. \tag{19.20}$$

Thus, from (19.16) it follows that

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{d} N\left(0, \left(\frac{\partial g^{**}(\theta_0)}{\partial \theta'}\mathbf{M}_*'(\theta_0)\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\right)^{-1}\right).$$

Hence we have

$$\xi_n^W \# \xi_n^S$$

as desired, where

$$\xi_n^S = \frac{1}{n}\hat{\lambda}_n'\frac{\partial g^{**}(\hat{\theta}_n^0)}{\partial \theta'}\mathbf{M}_*'(\hat{\theta}_n^0)\mathcal{I}^{-1}(\hat{\theta}_n^0)\mathbf{M}_*(\hat{\theta}_n^0)\frac{\partial g^{**\prime}(\hat{\theta}_n^0)}{\partial \theta}\hat{\lambda}_n. \tag{19.21}$$

Moreover, in view of (19.7) and the definition of the matrix $\mathbf{M}_*(\hat{\theta}_n^0)$ we have

$$\xi_n^S = \frac{1}{n}\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'}\mathbf{M}_*'(\hat{\theta}_n^0)\mathcal{I}^{-1}(\hat{\theta}_n^0)\mathbf{M}_*(\hat{\theta}_n^0)\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta}.$$

e) ASYMPTOTIC EQUIVALENCE BETWEEN ξ_n^L AND ξ_n^R:

As in the proof of Property 17.4, the Taylor series expansions of $L_n(\hat{\theta}_n^*)$ and $L_n(\hat{\theta}_n^0)$ around θ_0 imply

$$\begin{aligned}\xi_n^R &= 2\left(L_n(\hat{\theta}_n^*) - L_n(\hat{\theta}_n^0)\right)\\ &\# 2\frac{\partial L_n(\theta_0)}{\partial \theta'}(\hat{\theta}_n^* - \hat{\theta}_n^0)\\ &\quad -n(\hat{\theta}_n^* - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^* - \theta_0) + n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0).\end{aligned}$$

On the other hand, we have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\hat{\theta}_n^*)}{\partial \theta} \# \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^* - \theta_0).$$

Hence

$$\frac{\partial L_n(\theta_0)}{\partial \theta'}(\hat{\theta}_n^* - \hat{\theta}_n^0) \# \frac{\partial L_n(\hat{\theta}_n^*)}{\partial \theta'}(\hat{\theta}_n^* - \hat{\theta}_n^0) + n(\hat{\theta}_n^* - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^* - \hat{\theta}_n^0).$$

From (19.6) the first term in the right-hand side becomes

$$\begin{aligned}
\frac{\partial L_n(\hat{\theta}_n^*)}{\partial \theta'}(\hat{\theta}_n^* - \hat{\theta}_n^0) &= -\hat{\mu}_n' \frac{\partial g^*(\hat{\theta}_n^*)}{\partial \theta'}(\hat{\theta}_n^* - \hat{\theta}_n^0) \\
&\# -\frac{\hat{\mu}_n'}{\sqrt{n}} \frac{\partial g^*(\theta_0)}{\partial \theta'} \left(\sqrt{n}(\hat{\theta}_n^* - \theta_0) - \sqrt{n}(\hat{\theta}_n^0 - \theta_0)\right) \\
&\# -\frac{\hat{\mu}_n'}{\sqrt{n}} \sqrt{n} \left(g^*(\hat{\theta}_n^*) - g^*(\hat{\theta}_n^0)\right) \\
&\# 0
\end{aligned}$$

since $g^*(\hat{\theta}_n^*) = g^*(\hat{\theta}_n^0) = 0$.

Hence the likelihood ratio test statistic satisfies

$$\begin{aligned}
\xi_n^R \# & 2n(\hat{\theta}_n^* - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^* - \hat{\theta}_n^0) \\
& -n(\hat{\theta}_n^* - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^* - \theta_0) + n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0),
\end{aligned}$$

or equivalently

$$\xi_n^R \# n(\hat{\theta}_n^* - \hat{\theta}_n^0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^* - \hat{\theta}_n^0) \tag{19.22}$$

(see (17.21)).

Now, from (19.19) we have

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \# -\frac{1}{\sqrt{n}}\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\hat{\lambda}_n. \tag{19.23}$$

Hence, (19.21), (19.22), and (19.23) imply that $\xi_n^R \# \xi_n^S$ under H_0.

The remaining statements are proved as in Section 17.2. □

Computation of the test statistics ξ_n^W, ξ_n^S, and ξ_n^R can be cumbersome. The next two properties are then useful.

Property 19.3: *Under H_0, one obtains test statistics that are asymptotically equivalent to ξ_n^W, ξ_n^S, and ξ_n^R by replacing $\hat{\theta}_n^0$ by $\bar{\theta}_n^0$ or $\tilde{\theta}_n^0$ and $\hat{\theta}_n^*$ by $\tilde{\theta}_n^*$ in the expressions of these statistics given in Property 19.2, where*

$$\bar{\theta}_n^0 = \hat{\theta}_n^* - \mathcal{I}^{-1}(\hat{\theta}_n^*)\mathbf{M}_*(\hat{\theta}_n^*)\frac{\partial g^{**\prime}(\hat{\theta}_n^*)}{\partial \theta} \times \left(\frac{\partial g^{**}(\hat{\theta}_n^*)}{\partial \theta'}\mathbf{M}_*'(\hat{\theta}_n^*)\mathcal{I}^{-1}(\hat{\theta}_n^*)\mathbf{M}_*(\hat{\theta}_n^*)\frac{\partial g^{**\prime}(\hat{\theta}_n^*)}{\partial \theta}\right)^{-1} g^{**}(\hat{\theta}_n^*),$$

$$\tilde{\theta}_n^0 = \hat{\theta}_n - \mathcal{I}^{-1}(\hat{\theta}_n)\frac{\partial g'(\hat{\theta}_n)}{\partial \theta}\left(\frac{\partial g(\hat{\theta}_n)}{\partial \theta'}\mathcal{I}^{-1}(\hat{\theta}_n)\frac{\partial g'(\hat{\theta}_n)}{\partial \theta}\right)^{-1} g(\hat{\theta}_n),$$

$$\tilde{\theta}_n^* = \hat{\theta}_n - \mathcal{I}^{-1}(\hat{\theta}_n)\frac{\partial g^{*\prime}(\hat{\theta}_n)}{\partial \theta}\left(\frac{\partial g^{*}(\hat{\theta}_n)}{\partial \theta'}\mathcal{I}^{-1}(\hat{\theta}_n)\frac{\partial g^{*\prime}(\hat{\theta}_n)}{\partial \theta}\right)^{-1} g^{*}(\hat{\theta}_n),$$

and $\hat{\theta}_n$ is the unconstrained maximum likelihood estimator of θ.

PROOF:

a) First, we show that

$$\sqrt{n}(\bar{\theta}_n^0 - \hat{\theta}_n^0) \# 0, \quad \sqrt{n}(\tilde{\theta}_n^0 - \hat{\theta}_n^0) \# 0, \quad \sqrt{n}(\tilde{\theta}_n^* - \hat{\theta}_n^*) \# 0,$$

under H_0.

From (19.19) we have

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \# -\frac{1}{\sqrt{n}}\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\hat{\lambda}_n.$$

Then, using (19.20), we obtain

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0) \# \mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta} \times \left(\frac{\partial g^{**}(\theta_0)}{\partial \theta'}\mathbf{M}_*'(\theta_0)\mathcal{I}^{-1}(\theta_0)\mathbf{M}_*(\theta_0)\frac{\partial g^{**\prime}(\theta_0)}{\partial \theta}\right)^{-1}\sqrt{n}g^{**}(\hat{\theta}_n^*).$$

The latter asymptotic equivalence implies that

$$\sqrt{n}(\bar{\theta}_n^0 - \hat{\theta}_n^0) \# 0.$$

The other two asymptotic equivalences are established similarly by taking the unconstrained model as the maintained model.

b) Using a) it is then easy to verify that the substitutions indicated in Property 19.3 are allowed. For instance, using the second-order Taylor series expansions of $L_n(\hat{\theta}_n^0)$ and $L_n(\bar{\theta}_n^0)$ around θ_0, and taking their difference, we obtain

$$\begin{aligned} L_n(\hat{\theta}_n^0) - L_n(\bar{\theta}_n^0) \quad \# \quad & \frac{\partial L_n(\theta_0)}{\partial \theta}(\hat{\theta}_n^0 - \bar{\theta}_n^0) \\ & + \frac{n}{2}(\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) \\ & - \frac{n}{2}(\bar{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\bar{\theta}_n^0 - \theta_0). \end{aligned}$$

Thus

$$\begin{aligned} L_n(\hat{\theta}_n^0) - L_n(\bar{\theta}_n^0) \quad \# \quad & \frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_0)}{\partial \theta} \sqrt{n}(\hat{\theta}_n^0 - \bar{\theta}_n^0) \\ & + \frac{n}{2}(\hat{\theta}_n^0 - \bar{\theta}_n^0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \bar{\theta}_n^0) \\ & + n(\hat{\theta}_n^0 - \bar{\theta}_n^0) \mathcal{I}(\theta_0)(\bar{\theta}_n^0 - \theta_0). \end{aligned}$$

Since $\sqrt{n}(\hat{\theta}_n^0 - \bar{\theta}_n^0) \# 0$, we obtain $L_n(\hat{\theta}_n^0) - L_n(\bar{\theta}_n^0) \# 0$. Therefore

$$2\left(L_n(\hat{\theta}_n^*) - L_n(\bar{\theta}_n^0)\right) \# \xi_n^R.$$

Similarly, $\hat{\theta}_n^*$ can be replaced by $\tilde{\theta}_n^*$ in the expression of ξ_n^R. This can also be done in the expression of the Wald statistic ξ_n^W because

$$\sqrt{n}\left(g^{**}(\hat{\theta}_n^*) - g^{**}(\tilde{\theta}_n^*)\right) \# \sqrt{n} \frac{\partial g(\theta_0)}{\partial \theta'}(\hat{\theta}_n^* - \tilde{\theta}_n^*) \# 0.$$

Lastly, $\hat{\theta}_n^0$ can be replaced by $\bar{\theta}_n^0$ in the expression of ξ_n^S because

$$\frac{1}{\sqrt{n}}\left(\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta'} - \frac{\partial L_n(\bar{\theta}_n^0)}{\partial \theta}\right) \# \mathcal{I}(\theta_0)\sqrt{n}(\bar{\theta}_n^0 - \hat{\theta}_n^0) \# 0.$$

The same reasoning applies to $\tilde{\theta}_n^0$. □

Suppose that $g^*(\theta) = \theta_1 - \theta_1^0$, where θ_1 is a subvector of θ and θ_1^0 is a known vector. Let θ_2 denote the vector of remaining parameters of θ and $\tilde{\theta}_{2n}^*$ denote the corresponding subvector of $\tilde{\theta}_n^*$. We have

$$\tilde{\theta}_{2n}^* = \hat{\theta}_{2n} - \mathcal{I}^{21}(\hat{\theta}_n)(\mathcal{I}^{11}(\hat{\theta}_n))^{-1}(\hat{\theta}_{1n} - \theta_1^0). \tag{19.24}$$

Then, using the identity $\mathcal{I}^{21}(\mathcal{I}^{11})^{-1} = -\mathcal{I}_{22}^{-1}\mathcal{I}_{21}$, we obtain

$$\tilde{\theta}_{2n}^{*} = \hat{\theta}_{2n} + \mathcal{I}_{22}^{-1}(\hat{\theta}_n)\mathcal{I}_{21}(\hat{\theta}_n)(\hat{\theta}_{1n} - \theta_1^0). \tag{19.25}$$

Equation (19.24) shows that the approximation $\tilde{\theta}_{2n}^{*}$ of $\hat{\theta}_{2n}^{*}$ can be interpreted as the residuals in a regression of the unconstrained estimator of θ_2 on the unconstrained estimator of θ_1.

Property 19.3 can be used to construct various test statistics that are easy to compute and that are asymptotically equivalent under H_0 to the classical test statistics ξ_n^W, ξ_n^S, and ξ_n^R. Among these statistics, for instance, the statistic $2(L_n(\hat{\theta}_n^{*}) - L_n(\tilde{\theta}_n^0))$ can be computed from the unconstrained estimator $\hat{\theta}_n$ of θ only.

A second property can also simplify the required numerical computations.

Property 19.4: *Let* $(\xi_{n,*}^W, \xi_{n,*}^S, \xi_{n,*}^R)$ *and* $(\xi_{n,**}^W, \xi_{n,**}^S, \xi_{n,**}^R)$ *denote respectively the Wald, score, and likelihood ratio test statistics for testing*

$$\{g^{*}(\theta) = 0\} \quad \textit{and} \quad \{g^{*}(\theta) = 0 \textit{ and } g^{**}(\theta) = 0\}$$

against the unconstrained model. Then we have under H_0

$$\begin{aligned} \xi_n^W &\ \# \ \xi_{n,**}^W - \xi_{n,*}^W, \\ \xi_n^S &\ \# \ \xi_{n,**}^S - \xi_{n,*}^S, \\ \xi_n^R &\ \# \ \xi_{n,**}^R - \xi_{n,*}^R. \end{aligned}$$

PROOF: The third asymptotic equivalence is obvious since

$$\begin{aligned} \xi_n^R &= 2\left(L_n(\hat{\theta}_n^{*}) - L_n(\hat{\theta}_n^0)\right) \\ &= 2\left(L_n(\hat{\theta}_n^{*}) - L_n(\hat{\theta}_n) + L_n(\hat{\theta}_n) - L_n(\hat{\theta}_n^0)\right) \\ &= \xi_{n,**}^R - \xi_{n,*}^R. \end{aligned}$$

The other asymptotic equivalences follows at once from the asymptotic equivalences under H_0 of $\xi_{n,*}^R$, $\xi_{n,*}^S$, and $\xi_{n,*}^W$ on the one hand, and $\xi_{n,**}^R$, $\xi_{n,**}^S$, and $\xi_{n,**}^W$ on the other hand. □

For instance, for the Wald test statistic, the preceding property allows us to construct an asymptotically equivalent test statistic that depends on $\hat{\theta}_n$ only since $\xi_{n,**}^W$ and $\xi_{n,*}^W$ depend on this estimator only.

Lastly, we give an independence property, which is useful for the construction of sequential tests discussed in the next subsection.

Property 19.5: *The classical test statistics ξ_n^W, ξ_n^S, and ξ_n^R (or their asymptotically equivalent test statistics under H_0 defined in Property 19.4) are asymptotically independent of $\hat{\theta}_n^0$ under H_0.*

PROOF: Under H_0, we have

$$\xi_n^W \# \xi_n^S \# \xi_n^R \# n(\hat{\theta}_n^* - \hat{\theta}_n^0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^* - \hat{\theta}_n^0)$$

(see equation (19.22)). Hence, it suffices to show that the joint asymptotic distribution of $(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0), \sqrt{n}(\hat{\theta}_n^0 - \theta_0))$ is the product of the marginal asymptotic distributions.

From equation (19.14) we have

$$\sqrt{n}(\hat{\theta}_n^* - \theta_0)) \# \mathbf{M}_*'(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0).$$

Similarly, we have

$$\sqrt{n}(\hat{\theta}_n^0 - \theta_0)) \# \mathbf{M}'(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0),$$

where $\mathbf{M}(\theta_0)$ is an orthogonal projection matrix analogous to the matrix $\mathbf{M}_*(\theta_0)$ obtained by replacing g by g^*. Thus taking the difference between the above two equations gives

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0)) \# (\mathbf{M}_*'(\theta_0) - \mathbf{M}'(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0).$$

It follows that the vector $(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0), \sqrt{n}(\hat{\theta}_n^0 - \theta_0))$ is asymptotically normally distributed with a covariance matrix between $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n^0)$ and $\sqrt{n}(\hat{\theta}_n^0 - \theta_0)$ equal to

$$(\mathbf{M}_*'(\theta_0) - \mathbf{M}'(\theta_0))\mathcal{I}^{-1}(\theta_0)\mathbf{M}(\theta_0).$$

Since g^* is a subvector of g, the column space of $\mathbf{M}(\theta_0)$ is included in that of $\mathbf{M}_*(\theta_0)$. Therefore $\mathbf{M}(\theta_0) = \mathbf{M}(\theta_0)\mathbf{M}_*(\theta_0)$ and $\mathbf{M}'(\theta_0) = \mathbf{M}_*'(\theta_0)\mathbf{M}'(\theta_0)$. Hence

$$\begin{aligned}
&(\mathbf{M}_*'(\theta_0) - \mathbf{M}'(\theta_0))\mathcal{I}^{-1}(\theta_0)\mathbf{M}(\theta_0) \\
&\quad = \mathbf{M}_*'(\theta_0)(\mathbf{I} - \mathbf{M}'(\theta_0))\mathcal{I}^{-1}(\theta_0)\mathbf{M}(\theta_0) \\
&\quad = 0.
\end{aligned}$$

□

19.2.2 Sequential Tests of Nested Hypotheses in Implicit Form

a) The Testing Procedure

Here we want to generalize the results of Section 19.1 obtained for the linear model. We consider now a sequence of nested hypotheses defined by

$$H_0^k : \{g^{(K)}(\theta) = 0, g^{(K-1)}(\theta) = 0, \ldots, g^{(k)}(\theta) = 0\}, \quad k = 1, \ldots, K. \tag{19.26}$$

These hypotheses are such that $H_0^k \supset H_0^{k-1}$. The function $g^{(k)}$ is of dimension r_k. We let $g = (g^{(1)\prime} \ldots g^{(K)\prime})'$ and we assume that the Jacobian matrix $\partial g'/\partial\theta$ is of rank $r = \sum_{k=1}^{K} r_k$.

We consider the following testing procedure. First, the null hypothesis H_0^K is tested against the unconstrained model by means of a Wald, score, or likelihood ratio type test at level α_K. If H_0^K is rejected, the procedure ends. Otherwise, the null hypothesis H_0^{K-1} is tested against the model constrained by H_0^K. The procedure continues until either a null hypothesis H_0^k is rejected, in which case H_0^{k+1} is the most constraining hypothesis that is accepted, or until the null hypothesis H_0^1 is accepted.

The independence property that was established for the linear model remains valid asymptotically in the present context.

Property 19.6: *Let ξ_n^k be a Wald, score, or likelihood ratio test statistic (or any asymptotically equivalent test statistic) for testing H_0^k against the model constrained by H_0^{k+1}. By convention, H_0^{K+1} denotes the unconstrained model. Under H_0^k, the test statistics ξ_n^K, $\xi_n^{K-1}, \ldots, \xi_n^k$ are asymptotically independent. Moreover, ξ_n^i is asymptotically distributed as a $\chi^2(r_i)$, for every $i = k, \ldots, K$.*

PROOF: From Property 19.5 it follows that ξ_n^K is asymptotically independent of the maximum likelihood estimator $\hat{\theta}_n^K$ of θ under H_0^K. Moreover, applying Property 19.4 to the model constrained by H_0^K, it follows that the test statistics $\xi_n^{K-1}, \ldots, \xi_n^k$ are asymptotically equivalent under H_0^K to some test statistics that depend on the observations through $\hat{\theta}_n^K$ only. For instance, these statistics can be obtained as differences of Wald test statistics maintaining H_0^K. As a result, under H_0^k (which implies H_0^K), the test statistic ξ_n^K is asymptotically independent of $(\xi_n^{K-1}, \ldots, \xi_n^k)$.

Because H_0^k implies $H_0^{K-1}, \ldots, H_0^{k+1}$, a similar reasoning shows that, under H_0^k,

$$\begin{aligned} &\xi_n^{K-1} \quad \text{is asymptotically independent of} \quad (\xi_n^{K-2}, \ldots, \xi_n^k), \\ &\vdots \\ &\xi_n^{k+1} \quad \text{is asymptotically independent of} \quad \xi_n^k. \end{aligned}$$

Hence $\xi_n^K, \xi_n^{K-1}, \ldots, \xi_n^k$ are asymptotically independent as required, under H_0^k. The asymptotic chi-square distribution of $\xi_n^i, i = k, \ldots, K$, follows immediately from Property 19.2. □

b) Significance Level of the Testing Procedure

The preceding result allows us to compute the asymptotic levels of the sequential testing procedure defined above when a given hypothesis H_0^k is chosen as the null hypothesis.

Specifically, the critical region for testing H_0^k is

$$\bigcup_{i=k}^{K} W_n^i, \quad \text{with} \quad W_n^i = \{\xi_n^i > \chi^2_{1-\alpha_i}(r_i)\}.$$

Hence the asymptotic level associated with the above sequential testing procedure for testing H_0^k is

$$\lim_{n\to\infty} P_{H_0^k}\left(\bigcup_{i=k}^{K} W_n^i\right) = \lim_{n\to\infty}\left(1 - P_{H_0^k}\left(\bigcap_{i=k}^{K} {}^cW_n^i\right)\right).$$

Now, from the asymptotic independence given in Property 19.6 it follows that this asymptotic level is also equal to

$$1 - \prod_{i=k}^{K} \lim_{n\to\infty} P_{H_0^k}({}^cW_n^i),$$

i.e., to

$$1 - \prod_{i=k}^{K}(1 - \alpha_i). \tag{19.27}$$

Thus, as in Section 19.1.2, equation (19.27) can be used to choose appropriately the individual levels α_i so as to reach desired asymptotic

levels for testing H_0^k, $k = 1, \ldots, K$, provided these desired levels increase as k decrceases. Conversely, a set of given values for the α_i's imply some asymptotic levels for testing the hypotheses H_0^k that increase as k decreases. In particular, if $\alpha_i = \alpha$ for every i, then the asymptotic level for testing H_0^k is

$$1 - (1 - \alpha)^{K-k+1}.$$

If α is small, such an asymptotic level is approximately equal to

$$(K - k + 1)\alpha - \frac{(K - k + 1)(K - k)}{2}\alpha^2.$$

Thus, for instance, if $\alpha = 5\%$ and $K - k + 1 = 4$, then we obtain an asymptotic level equal approximately to 18.5%.

c) Consistency of the Sequential Testing Procedure

It is interesting to note that the above sequential testing procedure is consistent for testing separately the null hypothesis H_0^k. For the asymptotic power of the test is

$$\pi = \lim_{n\to\infty} P_{c_{H_0^k}} \left(\bigcup_{i=k}^{K} W_n^i \right) \geq \lim_{n\to\infty} P_{c_{H_0^k}} W_n^i, \quad \forall i = k, \ldots, K.$$

But every element θ belonging to ${}^cH_0^k$ also belongs to the alternative hypothesis of one of the tests based on $W_n^i, i = k, \ldots, K$, i.e., to one of the set defined by $H_0^{i+1} - H_0^i$, $i = k, \ldots, K$ (using the same convention as previously regarding H_0^{K+1}). Because each separate test is consistent, then $\pi = 1$. Hence the sequential procedure provides a consistent test of H_0^k.

19.2.3 Example

We consider the nonlinear regression model

$$Y_i = m(x_i, b) + u_i, \quad i = 1, \ldots, n,$$

where x_i is a vector of exogenous variables, b is a p-dimensional parameter vector, and $m(x_i, b)$ is a scalar function that is differentiable in b. The conditional distribution of $(u_i,\ i = 1, \ldots, n)$ given $(x_i,\ i = 1, \ldots, n)$ is the multivariate normal distribution $N(0, \sigma^2\mathbf{I})$, where σ^2 is an unknown parameter.

We want to test sequentially the hypotheses

$$H_0^k : \{g^{(K)}(b) = 0, \ldots, g^{(k)}(b) = 0\}, \quad k = K, \ldots, 1,$$

where $g^{(k)}$ is of dimension r_k. We let $g = (g^{(1)}, \ldots, g^{(K)})'$ and we assume that the matrix $\partial g'/\partial\theta$ is of rank $r = \sum_{k=1}^{K} r_k$.

Let $g_{(k)} = (g^{(K)}, \ldots, g^{(k)})'$. From (17.39) the Wald test statistic for testing the hypothesis H_0^k against the unconstrained model is

$$S_n^k = \frac{1}{\hat{\sigma}_n^2} g'_{(k)}(\hat{b}_n) \left(\frac{\partial g_{(k)}(\hat{b}_n)}{\partial b'} (\hat{\mathbf{M}}'\hat{\mathbf{M}})^{-1} \frac{\partial g'_{(k)}(\hat{b}_n)}{\partial b} \right)^{-1} g_{(k)}(\hat{b}_n),$$

where $\hat{b}_n$ and $\hat{\sigma}_n^2$ are the unconstrained maximum likelihood estimators of b and σ^2, and $\hat{\mathbf{M}}$ is the matrix whose generic term is $\partial m(x_i, \hat{b}_n)/\partial b_j$.

It follows from Property 19.4 that we can use

$$\xi_n^k = S_n^k - S_n^{k+1}$$

as a Wald test statistic for testing H_0^k against the model constrained by H_0^{k+1}. Such a statistic can easily be computed since it does not require any constrained maximization of the likelihood function.

Now, Property 19.6 implies that the test statistics $\xi_n^K, \ldots, \xi_n^k$ are asymptotically independent and distributed as $\chi^2(r_K), \ldots, \chi^2(r_k)$ respectively, under the hypothesis H_0^k.

19.3 Sequences of Null Hypotheses in Explicit Form: The General Case

As for the case of sequences of null hypotheses in implicit form studied in the previous section, we shall first generalize the results of Chapter 17 on Wald, score, and likelihood ratio tests to the case where the general or maintained hypothesis already imposes a constraint on the parameter θ. Then we shall apply these results to the testing of a sequence of null hypotheses in explicit form.

19.3.1 Test of a Hypothesis When the Maintained Hypothesis is Already Constrained

We consider the general framework of Section 19.2. But we now assume that the maintained model satisfies the constraint in the explicit form

$$\theta = h^*(\gamma^*), \quad \gamma^* \in \Gamma^* \subset I\!R^{q^*},$$

where h^* is a differentiable function from $\mathbb{R}^{q*}$ to $\mathbb{R}^p$.

Within such a general framework we want to test the null hypothesis H_0 in the explicit form

$$\gamma^* = h^{**}(\gamma), \quad \gamma \in \Gamma \subset \mathbb{R}^q,$$

where h^{**} is a differentiable function from $\mathbb{R}^q$ to $\mathbb{R}^{q*}$. In view of the maintained hypothesis, such a hypothesis can also be written as $\theta = h(\gamma)$, where $h = h^* \circ h^{**}$, where $\circ$ denotes the composition of functions.

The log-likelihood function of the model is $L_n^*(\gamma^*) = L_n(h^*(\gamma^*))$. Thus all the results of Section 17.4 apply directly. In particular, we can define the three test statistics ξ_n^W, ξ_n^S, and ξ_n^R. Moreover, these three test statistics are asymptotically equivalent under H_0.

To compute these test statistics, two properties can be useful. First, a property analogous to Property 19.4 can be easily proved. Such a property allows to compute the above test statistics as differences in the test statistic for testing H_0 against the unconstrained model and the test statistic for testing the maintained (constrained) model against the unconstrained model.

Second, let $\hat{\theta}_n^0 = h(\hat{\gamma}_n)$ denote the maximum likelihood estimator of θ under H_0. Let $\hat{\theta}_n^* = h^*(\hat{\gamma}_n^*)$ denote the maximum likelihood estimator of θ in the maintained model. In Section 17.4.4, we noted that

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n^0 - \bar{\theta}_n^0) &\# 0, \\ \sqrt{n}(\hat{\theta}_n^0 - \tilde{\theta}_n^0) &\# 0, \\ \sqrt{n}(\hat{\theta}_n^* - \tilde{\theta}_n^*) &\# 0, \end{aligned}$$

under H_0, where $\bar{\theta}_n^0 = h^{**}(\bar{\gamma}_n)$, $\tilde{\theta}_n^0 = h^{**}(\tilde{\gamma}_n)$, $\tilde{\theta}_n^* = h^*(\tilde{\gamma}_n^*)$, and $\bar{\gamma}_n$, $\tilde{\gamma}_n$, and $\tilde{\gamma}_n^*$ are obtained by minimizing respectively

$$\begin{aligned} &(\hat{\gamma}_n^* - h^{**}(\gamma))'(\hat{V}_{as}(\hat{\gamma}_n^*))^{-1}(\hat{\gamma}_n^* - h^{**}(\gamma)), \\ &(\hat{\theta}_n - h(\gamma))'(\hat{V}_{as}(\hat{\theta}_n))^{-1}(\hat{\theta}_n - h(\gamma)), \\ &(\hat{\theta}_n - h^*(\gamma^*))'(\hat{V}_{as}(\hat{\theta}_n))^{-1}(\hat{\theta}_n - h^*(\gamma^*)). \end{aligned}$$

Thus, as in Property 19.3, various test statistics that are asymptotically equivalent to ξ_n^W, ξ_n^S, and ξ_n^R under H_0 can be obtained by replacing $\hat{\theta}_n^0$ by $\bar{\theta}_n^0$ or or $\tilde{\theta}_n^0$, and $\hat{\theta}_n^*$ by $\tilde{\theta}_n^*$ in the expressions for ξ_n^W, ξ_n^S, and ξ_n^R.

19.3.2 Sequential Tests of Nested Hypotheses in Explicit Form

We consider now the sequence of null hypotheses in explicit form

$$H_0^k : \{\theta = h_K \circ h_{K-1} \circ \ldots \circ h_k(\gamma_k)\}, \quad k = K, \ldots, 1.$$

In this case we have an asymptotic independence property that is analogous to Property 19.6. In particular, let ξ_n^k be a Wald, score or likelihood ratio test statistic for testing H_0^k against the model constrained by H_0^{k+1} (or a test statistic that is asymptotically equivalent under H_0^k). Then the test statistics $\xi_n^K, \ldots, \xi_n^k$ are asymptotically independent under H_0^k.

19.3.3 Example: Tests for a Common Root

a) We consider the following dynamic model

$$Y_t = \theta_1 Y_{t-1} + \theta_2 x_t + \theta_3 x_{t-1} + u_t, \quad t = 1, \ldots T,$$

where Y_0 is considered as fixed and the u_t's are independent and identically distributed $N(0, \theta_4)$. Introducing the lag operator L, this equation can also be written as

$$(1 - \theta_1 L) Y_t = (\theta_2 + \theta_3 L) x_t + u_t, \quad t = 1, \ldots, T.$$

The hypothesis that the polynomials $1 - \theta_1 L$ and $\theta_2 + \theta_3 L$ have a common root can be written in explicit form as

$$\begin{cases} 1 - \theta_1 L &= 1 - \gamma_1^* L, \\ \theta_2 + \theta_3 L &= \gamma_2^* (1 - \gamma_1^* L), \\ \theta_4 &= \gamma_3^*, \end{cases}$$

where $\theta_1 = \gamma_1^*, \theta_2 = \gamma_2^*, \theta_3 = -\gamma_1^* \gamma_2^*$ and $\theta_4 = \gamma_3^*$.

Let $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, and $\hat{\theta}_4$ denote the OLS estimators of θ_1, θ_2, θ_3, and θ_4 respectively. We know that $\hat{\theta}_4$ is asymptotically independent of $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$. Hence a possible test statistic for testing the hypothesis of a common root is

$$\begin{aligned} \xi_T^* = \min_{\gamma_1^*, \gamma_2^*} \quad & (\hat{\theta}_1 - \gamma_1^*, \hat{\theta}_2 - \gamma_2^*, \hat{\theta}_3 + \gamma_1^* \gamma_2^*)' \mathbf{X}' \mathbf{X} \\ & \times (\hat{\theta}_1 - \gamma_1^*, \hat{\theta}_2 - \gamma_2^*, \hat{\theta}_3 + \gamma_1^* \gamma_2^*) \end{aligned}$$

(see Remark 17.10), where $\mathbf{X}$ is the $(T \times 3)$ matrix whose t-th row is (Y_{t-1}, x_t, x_{t-1}). This test statistic is asymptotically equivalent to the

Wald, score, and likelihood ratio test statistics. The corresponding test is the test with critical region $\{\xi_T^* \geq \chi^2_{1-\alpha}(1)\}$.

b) If the hypothesis of common root is accepted, we may want to test under this hypothesis, the additional restriction that the two preceding polynomials are equal, i.e.

$$\begin{cases} \gamma_1^* &= \gamma_1, \\ \gamma_2^* &= 1, \\ \gamma_3^* &= \gamma_2. \end{cases}$$

In the unconstrained model this hypothesis is

$$\begin{cases} \theta_1 &= \gamma_1, \\ \theta_2 &= 1, \\ \theta_3 &= -\gamma_1, \\ \theta_4 &= \gamma_2. \end{cases}$$

Thus a test statistic of this hypothesis in the unconstrained model is

$$\xi_T^{**} = \min_{\gamma_1} (\hat{\theta}_1 - \gamma_1, \hat{\theta}_2 - 1, \hat{\theta}_3 + \gamma_1)' \mathbf{X}'\mathbf{X} (\hat{\theta}_1 - \gamma_1, \hat{\theta}_2 - 1, \hat{\theta}_3 + \gamma_1),$$

which is asymptotically distributed as a $\chi^2(2)$ distribution under the null hypothesis of equality of the two polynomials.

To test the hypothesis of equality of the two polynomials in the model constrained by the hypothesis of a common root, we can apply the general results established previously. For instance, we can use the test statistic

$$\xi_T = \xi_T^{**} - \xi_T^*.$$

Then, under the combined null hypothesis of a common root and equality of the two polynomials, the test statistic ξ_T follows a $\chi^2(1)$ distribution. The critical region of the corresponding test is $\{\xi_T^* \geq \chi^2_{1-\alpha}(1)\}$. Moreover, under this combined null hypothesis, the test statistic ξ_T is asymptotically independent of the statistic ξ_T^* used for testing the hypothesis of a common root only.

19.4 General Multiple Tests

As in the previous sections we consider a conditional parametric model. However, we shall no longer assume that the null hypotheses of interest are nested. Specifically, we shall consider two types of situations where

this is not the case. In the first type of situations the model is given and we are interested in a set of nonnested null hypotheses in such a model. In the second type of situations the model is considered as a null hypothesis and we want to test this null hypothesis against more general models. As indicated in Section 14.1, the tests corresponding to the first type of situations are sometimes called *significance tests*, while the tests corresponding to the second type of situations are called *specification tests.*

19.4.1 Tests of Nonnested Null Hypotheses in a Given Model

We consider a conditional parametric model with a (conditional) log-likelihood function $L_n(\theta)$. The null hypotheses of interest are

$$H_0^{(k)} : \{g^{(k)}(\theta) = 0\}, \quad k = 1, \ldots, K.$$

We shall be especially interested in the hypothesis H_0, which is the intersection of all the hypotheses $H_0^{(k)}$, i.e.

$$H_0 : \{g(\theta) = 0\} = \{g^{(1)}(\theta) = 0, \ldots, g^{(K)}(\theta) = 0\}.$$

To test the hypothesis H_0, it is natural to use one of the testing procedures described in Chapter 17. When the hypothesis H_0 is rejected, however, such a testing procedure does not indicate which of the hypotheses $H_0^{(k)}$ are not satisfied and hence which of these separate hypotheses are responsible for the rejection of H_0.

To gain some information about the source of this rejection, one can use two different general testing procedures: One is the *Bonferroni procedure* and the other is *Scheffé procedure.*

a) The Bonferroni Procedure

In the Bonferroni procedure one tests the hypothesis $H_0^{(k)}$ at a given level α_k against the unconstrained model by means of one of the classical asymptotic methods. Let $W_n^{(k)} = \{\xi_n^{(k)} \geq \chi^2_{1-\alpha_k}(r_k)\}$ be the critical region of this test. Then the hypothesis H_0 is accepted if all the hypotheses $H_0^{(k)}$ are accepted.

Unlike in the preceding sections, the critical regions $W_n^{(k)}$ are, in general, not asymptotically independent under H_0. As a result, it is

difficult to evaluate exactly the significance level for testing H_0. The next property, however, can provide a useful bound.

Property 19.7: *In the Bonferroni procedure the test of H_0 has an asymptotic significance level that is smaller or equal to $\sum_{k=1}^{K} \alpha_k$. Moreover, the test is consistent.*

PROOF: First, the critical region associated with the test of H_0 is

$$W_n = \bigcup_{k=1}^{K} W_n^{(k)}.$$

Hence, Bonferroni inequality implies that

$$\lim_{n\to\infty} P_{H_0}(W_n) \leq \lim_{n\to\infty} \sum_{k=1}^{K} P_{H_0}(W_n^{(k)}) = \sum_{k=1}^{K} \alpha_k.$$

Second, if θ does not satisfy H_0, then θ does not satisfy at least one of the hypotheses $H_0^{(k)}$. Thus we have

$$\lim_{n\to\infty} P_\theta(W_n) \geq \lim_{n\to\infty} P_\theta(W_n^{(k)}) = 1$$

because $W_n^{(k)}$ is the critical region of a consistent test for $H_0^{(k)}$. Hence $\lim_{n\to\infty} P_\theta(W_n) = 1$ for every $\theta \in {}^cH_0$, i.e., the test of H_0 in the Bonferroni procedure is consistent. □

In particular, supose that all the α_k's are equal to α (say). Then, it follows from Property 19.7 that the asymptotic level of the test of H_0 in the Bonferroni procedure is bounded above by $K\alpha$. Now, in the preceding sections the asymptotic level was $1-(1-\alpha)^K$, which differs from the upper bound $K\alpha$ by $K(K+1)/2 \times \alpha^2$ approximately. Thus, if $\alpha = 5\%$ and $K = 4$, the Bonferroni upper bound is 20% while the exact asymptotic level in case of asymptotic independence is 18.5%. In this example the possible increase in the asymptotic significance level due to the loss of asymptotic independence is not substantial. This may not always be the case. For instance, if K is large, in particular, if $K\alpha$ is larger than 1, the Bonferroni bound is no longer informative.

b) The Scheffé Procedure

Consider the Wald test of the null hypothesis H_0: $\{g(\theta) = 0\}$. The Wald test statistic is

$$\xi_n^W = n g'(\hat{\theta}_n) \hat{\Omega}_n^{-1} g(\hat{\theta}_n), \tag{19.28}$$

where

$$\hat{\Omega}_n = \frac{\partial g(\hat{\theta}_n)}{\partial \theta'} \mathcal{I}^{-1}(\hat{\theta}_n) \frac{\partial g'(\hat{\theta}_n)}{\partial \theta}.$$

The critical region of this test is

$$\{\xi_n^{(W)} \geq \chi^2_{1-a}(r)\} \quad \text{with} \quad r = \sum_{k=1}^{K} r_k.$$

Consider now a null hypothesis of the form $H_0(a)$: $\{a'g(\theta) = 0\}$, where a is a given vector of $\mathbb{R}^r$. Then the corresponding Wald test statistic is

$$\xi_n^W(a) = \frac{n(a'g'(\hat{\theta}_n))^2}{a'\hat{\Omega}_n a}. \tag{19.29}$$

This statistic follows a $\chi^2(1)$ distribution under $H_0(a)$.

Since H_0 implies $H_0(a)$ for every a, then all the test statistics $\xi_n^W(a)$ are asymptotically $\chi^2(1)$ distributed under H_0. The next property allows us to determine the asymptotic distribution of $\max_a \xi_n^W(a)$.

Property 19.8: *We have* $\max_a \xi_n^W(a) = \xi_n^W$. *Moreover, the maximum is attained when* $a = \hat{\Omega}_n^{-1} g(\hat{\theta}_n)$.

PROOF : The definition of the test statistic $\xi_n^W(a)$ implies that $\xi_n^W(a)$ is invariant with respect to multiplication of a by a scalar. Hence, without loss of generality, we can impose a normalization condition on a. A convenient one is $a'\hat{\Omega}_n a = 1$.

Given such a normalization condition, the maximization of $\xi_n^W(a)$ reduces to

$$\begin{cases} \max_a (a'g(\hat{\theta}_n))^2 \\ \\ \text{subject to} \quad a'\hat{\Omega}_n a = 1. \end{cases}$$

This maximization problem can also be written as

$$\begin{cases} \max_a (a'\hat{\Omega}_n^{1/2}\hat{\Omega}_n^{-1/2} g(\hat{\theta}_n))^2 \\ \\ \text{subject to} \quad a'\hat{\Omega}_n a = 1. \end{cases}$$

Then, from Cauchy–Schwarz inequality we have

$$\begin{aligned} (a'\hat{\Omega}_n^{1/2}\hat{\Omega}_n^{-1/2} g(\hat{\theta}_n))^2 &\leq a'\hat{\Omega}_n a g'(\hat{\theta}_n)\hat{\Omega}_n^{-1} g(\hat{\theta}_n) \\ &= g'(\hat{\theta}_n)\hat{\Omega}_n^{-1} g(\hat{\theta}_n). \end{aligned}$$

Moreover, the maximum is attained when

$$a = \frac{\hat{\Omega}_n^{-1} g(\hat{\theta}_n)}{(g'(\hat{\theta}_n)\hat{\Omega}_n^{-1} g(\hat{\theta}_n))^{1/2}}, \tag{19.30}$$

which satisfies the constraint $a'\hat{\Omega}_n a = 1$, as required. Thus the maximal value of $\xi_n^W(a)$ is $ng'(\hat{\theta}_n)\hat{\Omega}_n^{-1} g(\hat{\theta}_n) = \xi_n^W$. □

Property 19.8 implies that the statistic $\max_a \xi_n^W(a)$ is asymptotically $\chi^2(r)$ distributed under H_0 and that

$$\{\xi_n^{(W)} < \chi^2_{1-\alpha}(r)\} = \{\xi_n^{(W)}(a) < \chi^2_{1-\alpha}(r), \forall a\}.$$

Thus, when using a Wald test based on ξ_n^W, the hypothesis H_0 is accepted if and only if

$$\xi_n^W(a) < \chi^2_{1-\alpha}(r) \tag{19.31}$$

for every vector a. Conversely, if H_0 is rejected, then some of the inequalities in (19.31) must be violated. In particular, the inequality corresponding to the optimal a given in Property 19.8 cannot be satisfied.

Scheffé's procedure thus consists in performing a global test of H_0 by means of the test statistic ξ_n^W. If H_0 is rejected, then one can determine the linear forms $a'g$ that are responsible for this rejection, i.e., the linear forms associated with the vectors a that do not satisfy (19.31). In particular, one may single out some of the components of g that are responsible for the rejection of H_0.

19.4.2 Tests of a Given Null Hypothesis Against Different Models

The conditional parametric model defined by the conditional densities $f(y_i \mid x_i; \theta)$ is now taken as the null hypothesis, which is tested against K larger models. It is assumed that these larger models are all parametric and that H_0 can be obtained by imposing some restrictions on the parameters of the larger models. Again we can use the three classical asymptotic tests. In the present case the score test turns out to be frequently the most useful because it requires only the maximum likelihood estimator under H_0, which remains constant.

The null hypothesis H_0 is accepted if the tests against the various alternatives all lead to accepting H_0. As in Section 19.4.1, Bonferroni

inequality can be used to show that the asymptotic level of such a multiple testing procedure for testing H_0 is bounded above by $\sum_{k=1}^{K} \alpha_k$, where α_k is the asymptotic level of the k-th individual test. In this multiple testing procedure the general or maintained model is not explicitly defined. However, if the maintained model is defined as the union of the various alternative models, the preceding multiple procedure leads to a consistent test of H_0 for the same reasons as those given in Property 19.7.

Lastly, as the next examples show, the different test statistics are sometimes asymptotically independent under H_0. In these favorable cases it is then possible to compute exactly the asymptotic level of the test for H_0 by the formula

$$1 - \prod_{k=1}^{K} (1 - \alpha_k).$$

Example 19.1: Omitted Variables

We consider the linear model

$$y = \mathbf{X}b + u, \quad u \sim N(0, \sigma^2 \mathbf{I}),$$

where $\mathbf{X}$ is a $(n \times h)$ matrix of rank h.

This model is considered as the null hypothesis H_0. Let $\hat{b}_0$ denote the maximum likelihood estimator of b (i.e., the OLS estimator). Let $\hat{u}_0 = y - \mathbf{X}\hat{b}_0$ denote the vector of residuals.

One wants to test whether an additional explanatory variable x_1 is relevant. Let X_1 denote the vector of observations on x_1. Let $\mathbf{X}_1 = (\mathbf{X}, X_1)$. Then the alternative model is

$$y = \mathbf{X}b + X_1 b_1 + u.$$

The null hypothesis is H_0: $b_1 = 0$.

From (17.38) the critical region of the score test is $\{\xi_{n,1}^S \geq \chi_{1-\alpha}^2(1)\}$ with

$$\begin{aligned} \xi_{n,1}^S &= \frac{1}{\hat{\sigma}_0^2} \hat{u}_0' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1 \hat{u}_0 \\ &= \frac{1}{\hat{\sigma}_0^2} \|\mathbf{M}_1 \hat{u}_0\|^2, \end{aligned}$$

where $\mathbf{M}_1$ is the matrix of the orthogonal projection on the subspace generated by the columns of $\mathbf{X}_1$.

In particular, if the vector X_1 is orthogonal to the columns of $\mathbf{X}$, the expression for the score test simplifies. Let $\mathbf{M}$ denote the matrix of the orthogonal projection on to the space generated by the columns of $\mathbf{X}$. Then

$$\mathbf{M}_1\hat{u}_0 = \mathbf{M}\hat{u}_0 + \frac{X_1X_1'\hat{u}_0}{X_1'X_1} = \frac{X_1X_1'\hat{u}_0}{X_1'X_1}$$

because $\hat{u}_0$ is orthogonal to the columns of $\mathbf{X}$. Hence

$$\xi_{n,1}^S = \frac{1}{\hat{\sigma}_0^2}\frac{(X_1'\hat{u}_0)^2}{X_1'X_1}.$$

Let X_2 be another column vector that is orthogonal to the column space of $\mathbf{X}$. Suppose that one also wishes to test whether X_2 is relevant. Then the corresponding test statistic is

$$\xi_{n,2}^S = \frac{1}{\hat{\sigma}_0^2}\frac{(X_2'\hat{u}_0)^2}{X_2'X_2}.$$

Now, if X_1 and X_2 are orthogonal to each other, then the statistics $X_1'\hat{u}_0$ and $X_2'\hat{u}_0$ are mutually independent under the null hypothesis because these statistics are normally distributed with a covariance equal to

$$X_1'V(\hat{u}_0)X_2 = \sigma^2X_1'(\mathbf{I}-\mathbf{M})X_2 = \sigma^2X_1'X_2 = 0.$$

It follows that the score test statistics $\xi_{n,1}^S$ and $\xi_{n,2}^S$ are asymptotically independent under the null hypothesis.

Example 19.2: We consider the linear model

$$Y_t = x_t'b + u_t,$$

where the explanatory variables x_t' do not include any lagged endogenous variables.

(i) A Test for Autocorrelation

The null hypothesis H_0 is that the errors u_t's are independently and identically distributed as $N(0,\sigma^2)$.

First, we consider testing H_0 in the more general model where

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1,$$

and the ε_t's are independently and identically distributed as $N(0,\sigma^2)$. Thus the null hypothesis H_0 corresponds to the constraint $\rho = 0$.

The test of such a zero correlation among the errors was studied in Section 17.1.4. In particular, the critical region of the score test at level α_1 of such a hypothesis is

$$\left\{\xi_T^{(1)} \geq \chi^2_{1-\alpha_1}(1)\right\},$$

where $\xi_T^{(1)} = T\hat{r}^2$ and $\hat{r}$ is the sample correlation coefficient between the residuals $\hat{u}_t$ and $\hat{u}_{t-1}$ obtained by ordinary least squares with a constant in the regression.

(ii) A Test for Heteroscedasticity

We now consider the same linear model and we want to test the same null hypothesis but against a model that displays some heteroscedasticity. Specifically, in the alternative model the u_t's are independent and distributed as

$$u_t \sim N(0, a_1 + a_2 z_t),$$

where the z_t's are positive exogenous variables, and a_1 and a_2 are positive parameters. Thus the null hypothesis H_0 corresponds to $a_2 = 0$.

It is easy to show that the elements corresponding to the pairs (b_i, a_j) in the Fisher information matrix $\mathcal{I}$ of the heteroscedastic model are null. Therefore the score test statistic is

$$\xi_T^{(2)} = \frac{1}{T}\left(\frac{\partial L_T}{\partial a}\right)'_{H_0} \hat{\mathcal{I}}_{aa}^{-1} \left(\frac{\partial L_T}{\partial a}\right)'_{H_0},$$

where $a' = (a_1, a_2)$, $\mathcal{I}_{aa}$ is the (2×2) submatrix of $\mathcal{I}$ corresponding to a, and

$$L_T = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log(a_1 + a_2 z_t) - \frac{1}{2}\sum_{t=1}^{T}\frac{(y_t - x_t b)^2}{a_1 + a_2 z_t}.$$

As mentioned in Chapter 17, we can let

$$T\hat{\mathcal{I}}_{aa} = \left(-E\frac{\partial^2 L_T}{\partial a \partial a'}\right)_{H_0}.$$

In the present case the latter matrix is

$$\left(E\frac{\partial^2 L_T}{\partial a \partial a'}\right)_{H_0} = -\frac{1}{2\hat{\sigma}^4}\sum_{t=1}^{T}\begin{pmatrix} 1 & z_t \\ z_t & z_t^2 \end{pmatrix},$$

where $\hat{\sigma}^2$ denotes the estimator of a_1 under H_0. That is, $\hat{\sigma}^2$ is the OLS estimator of the error variance.

On the other hand, we have

$$\left(\frac{\partial L_T}{\partial a}\right)_{H_0} = \frac{1}{2\hat{\sigma}^4}\begin{pmatrix} 0 \\ \sum_{t=1}^T (z_t - \bar{z})\hat{u}_t^2 \end{pmatrix}.$$

Hence the score test statistic is

$$\xi_T^{(2)} = \frac{1}{2\hat{\sigma}^4}\frac{\left(\sum_{t=1}^T (z_t - \bar{z})\hat{u}_t^2\right)^2}{\sum_{t=1}^T (z_t - \bar{z})^2}.$$

Let $\hat{a}_2$ denote the OLS estimator of a_2 in the auxiliary regression

$$\hat{u}_t^2 = a_1 + a_2 z_t + v_t, \tag{19.32}$$

where $Ev_t = 0$ and $V(v_t) = 2\hat{\sigma}^4$. We obtain

$$\xi_T^{(2)} = \frac{\hat{a}_2^2}{V(\hat{a}_2)}$$

since

$$\hat{a}_2 = \frac{\sum_{t=1}^T (z_t - \bar{z})\hat{u}_t^2}{\sum_{t=1}^T (z_t - \bar{z})^2},$$

$$V(\hat{a}_2) = \frac{2\hat{\sigma}^4}{\sum_{t=1}^T (z_t - \bar{z})^2}.$$

It follows that $\xi_T^{(2)}$ is asymptotically equivalent under H_0 to the square of the Student statistic for testing $a_2 = 0$ in the auxiliary regression model (19.32). See Exercise 19.4.

The critical region of the score test at level α_2 is

$$\left\{\xi_T^{(2)} \geq \chi^2_{1-\alpha_2}(1)\right\}.$$

(iii) A Test for Autocorrelation or Heteroscedasticity

Suppose now that we wish to test the null hypothesis H_0 in a model that displays simultaneously some autocorrelation and heteroscedasticity. Specifically, we consider the more general model defined by

$$\begin{cases} Y_t &= x_t'b + u_t, \\ u_t &= \rho u_{t-1} + \varepsilon_t, \\ \varepsilon_t &\sim N(0, a_1 + a_2 z_t). \end{cases} \tag{19.33}$$

It is easy to show (see Exercise 19.5) that the information matrix of this model has the following structure:

$$\begin{array}{cc} & \begin{array}{ccc} b & a & \rho \end{array} \\ \begin{array}{c} b \\ a \\ \rho \end{array} & \left\{\begin{array}{ccc} \times & 0 & 0 \\ 0 & \times & 0 \\ 0 & 0 & \times \end{array}\right. \end{array} \tag{19.34}$$

where a "×" indicates a nonzero element. It follows that the score test for testing simultaneously the absence of autocorrelation and heteroscedasticity can be based on the test statistic

$$\xi_T = \xi_T^{(1)} + \xi_T^{(2)},$$

where ξ_T^1 and ξ_T^2 are asymptotically independent under H_0. For an asymptotic level equal to α the critical region of the joint test is

$$\left\{\xi_T \geq \chi^2_{1-\alpha}(2)\right\}.$$

On the other hand, the test of H_0 defined by the critical region

$$W_T = \left\{\xi_T^{(1)} \geq \chi^2_{1-\alpha_1}(1)\right\} \cup \left\{\xi_T^{(2)} \geq \chi^2_{1-\alpha_2}(1)\right\}$$

has an asymptotic level equal to $1-(1-\alpha_1)(1-\alpha_2) = \alpha_1 + \alpha_2 - \alpha_1\alpha_2$. Note that such a test is consistent not only against the model that is the union of the two component alternative models used in the separate tests, but also against the more general model (19.33) that allows simultaneously the presence of some autocorrelation and heteroscedasticity. For, if the true value of the parameter of the model (19.33) does not satisfy the null hypothesis H_0, then ξ_T converges to $+\infty$. Hence, $\xi_T^{(1)}$ or $\xi_T^{(2)}$ must converge to $+\infty$. Therefore $P(W_T)$ converges to one.

19.5 Exercises

EXERCISE 19.1: Consider the framework of Section 18.1. Assume that $\mathcal{I}_0 = \mathcal{J}_0$. Then, using the expressions (18.11) and (18.15) for $\sqrt{n}(\hat{\theta}_n^0 - \theta_0)$ and $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0)$ respectively, show that these two statistics are asymptotically independent under H_0.

EXERCISE 19.2: Consider the framework of Section 18.1. Using Exercise 19.1 and assuming $\mathcal{I}_0 = \mathcal{J}_0$, show that an asymptotic independence property analogous to Property 19.6 can be established.

EXERCISE 19.3: Consider the linear model

$$y = \mathbf{X}b + u, \quad u \sim N(0, \sigma^2\mathbf{I}).$$

Let $\mathbf{M}$ denote the matrix of the orthogonal projection on to the subspace generated by the columns of $\mathbf{X}$. Using the injectivity property of the Laplace transform (see Property B.12), show that the pair of statistics $(\mathbf{M}Y, \|(\mathbf{I} - \mathbf{M})Y\|^2)$ is sufficient and complete.

EXERCISE 19.4: Let U be a random variable distributed as $N(0, \sigma^2)$. Show that the variance of U^2 is $2\sigma^4$. Conclude that the sample variance of $\hat{u}_t^2$ in the auxiliary regression (19.32) has the same limit as $2\hat{\sigma}^4$, namely $2\sigma^4$.

EXERCISE 19.5: Using the form of the likelihood function of the model (19.33), show that the Fisher information matrix of this model has the structure indicated in (19.34).

EXERCISE 19.6: Consider an heteroscedastic model analogous to the one considered in Example 19.2. Assume, however, that the u_t's are distributed as $N(0, h(a_1 + a_2 z_t))$ instead of $N(0, a_1 + a_2 z_t)$, where h is a differentiable positive monotone function. Show that the score test statistic of the null hypothesis $a_2 = 0$ does not depend on the function h and hence, is identical to the score test statistic obtained in Example 19.2.

EXERCISE 19.7: Generalize Example 19.1 by considering tests for omission of sets of explanatory variables. In particular, characterize the cases for which the test statistics are asymptotically independent under the null hypothesis.

EXERCISE 19.8: Consider the example of Section 19.3.3 and write the hypotheses defined there in implicit form. Then develop a testing procedure based on Wald type tests.

19.6 References

Savin, N.E. (1980). "The Bonferroni and Scheffé Multiple Comparison Procedures," *Review of Economic Studies*, 47, 255–273.

Savin, N.E. (1984). "Multiple Hypotheses Testing," in *Handbook of Econometrics*, Vol. 2, North-Holland.

Scheffé, H. (1950). *The Analysis of Variance*, Wiley.

Scheffé, H. (1953). "A Method of Judging all Contrasts in the Analysis of Variance," *Biometrika*, 40, 87–104.

CHAPTER 20

Set Estimation and Confidence Regions

We consider a parametric statistical model $(\mathcal{Y}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$. We are again interested in a function g of the parameter vector θ, where $g(\theta) \in G = g(\Theta) \subset \mathbb{R}^q$. Unlike in previous chapters, however, we want here to construct a subset of G that is "likely" to contain the true value of g, i.e., $g(\theta)$ when θ is the true value of the parameter vector. In what follows we shall define the precise sense to be given to the qualification "likely." Such a definition depends on whether a classical approach (Sections 20.1 to 20.7) or a Bayesian approach (Section 20.8) is adopted.

20.1 Classical Approach: Definitions

Within a classical approach, we want to find a random subset $S(Y)$ of G that contains $g(\theta)$ with a given probability $1-\alpha$ and which is independent of θ. Formally, we have the following definition.

Definition 20.1: *A confidence region at (confidence) level* $1-\alpha$ *for* $g(\theta)$ *is a family of subset of* $G = g(\theta)$ *indexed by* $y \in \mathcal{Y}$, *denoted* $\{S(y); y \in \mathcal{Y}\}$, *such that*

$$\forall \theta \in \Theta : P_\theta(y : S(y) \ni g(\theta)) = 1 - \alpha.$$

The latter equality is also written as

$$\forall \theta \in \Theta : P_\theta(S(Y) \ni g(\theta)) = 1 - \alpha.$$

It is important to emphasize that a confidence region is a *random* subset of G since it depends on $y \in \mathcal{Y}$, which is random. *Ex ante*, i.e.,

before observing the outcome y of the random experiment, this random subset has a probability $1-\alpha$ of containing the value $g(\theta)$ for each $\theta \in \Theta$. However, once the outcome y is observed, the subset $S(y)$ is deterministic. As a result, the quantity $P_\theta(S(y) \ni g(\theta))$ is equal to either 1 or 0 according to whether $g(\theta)$ belongs or does not belong to $S(y)$. Thus, given the observed value y, it is not correct to write that $P_\theta(S(y) \ni g(\theta))$ is equal to $1-\alpha$.

An important concept that is useful for finding confidence regions is the concept of *pivotal function.*

Definition 20.2: *A pivotal function for $g(\theta)$ is a function ν defined on $\mathcal{Y} \times G$ such that the probability distribution of $\nu(Y, g(\theta))$ is independent of $\theta \in \Theta$.*

Note that pivotal functions and ancillary statistics share the same property, which is to have a probability distribution that is independent of θ. However, unlike ancillary statistics, a pivotal function has $g(\theta)$ as one of its arguments by definition. As a result, $\nu(Y, g(\theta))$ is not a statistic.

Note also that the probability distribution of $\nu(Y, g(\theta))$ depends on θ in two ways: (i) Explicitly through $g(\theta)$, which appears as one of its arguments, and (ii) implicitly through Y whose probability distribution is P_θ. Thus the function ν is pivotal if these two effects cancel out.

Now, a pivotal function is used as follows. Let Π_ν denote the probability distribution of $\nu(Y, g(\theta))$. By definition, such a probability distribution is independent of θ. Let B be a subset of values for ν such that $\Pi_\nu(B) = 1-\alpha$. Then

$$\forall \theta \in \Theta,\ P_\theta(\nu(Y, g(\theta)) \in B) = 1-\alpha.$$

Let $S(y) = \{d \in g(\Theta) : \nu(y, d) \in B\}$. Then we obtain at once

$$\forall \theta \in \Theta,\ P_\theta(S(Y) \ni g(\theta)) = 1-\alpha.$$

Thus the subset S defines a confidence region for $g(\theta)$ at level $1-\alpha$.

20.2 Examples

20.2.1 Interval Estimation of the Mean of a Normal Distribution With Known Variance

We consider the sampling model $(\mathbb{R}, N(\theta, \sigma_0^2))^{\otimes n}$, where $\theta \in \mathbb{R}$ and σ_0 is known. We seek a confidence interval for θ.

The function $\sqrt{n}(\bar{Y} - \theta)/\sigma_0$ is pivotal for θ because its probability distribution is $N(0,1)$ and hence independent of θ. Thus let u denote the $(1 - \alpha/2)$-th quantile of the $N(0,1)$ distribution. We have

$$\forall \theta,\ P_\theta\left(-u \leq \sqrt{n}\frac{\bar{Y} - \theta}{\sigma_0} \leq u\right) = 1 - \alpha,$$

i.e.

$$\forall \theta,\ P_\theta\left(\bar{Y} - \frac{u\sigma_0}{\sqrt{n}} \leq \theta \leq \bar{Y} + \frac{u\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

Therefore the interval defined by the boundaries $\bar{Y} \pm u\sigma_0/\sqrt{n}$ is a confidence interval at level $1 - \alpha$ for θ.

20.2.2 Interval Estimation of the Mean of a Normal Distribution With Unknown Variance

We consider the sampling model $(\mathbb{R}, N(\mu, \sigma^2))^{\otimes n}$, where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ are now both unknown. We seek a confidence interval for the mean μ. We let $\theta = (\mu, \sigma^2)$. The function g corresponds here to the first component of θ, i.e., $g(\theta) = \mu$.

The function $\sqrt{n}(\bar{Y} - \mu)/\sigma$ is no longer pivotal for μ. Although its probability distribution is independent of θ, this function does not depend on θ through μ only. From Theorem B.7 and Definition B.22, however, we know that the function $\sqrt{n}(\bar{Y} - \mu)/s$, where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

has a fixed distribution, namely, the Student distribution with $n - 1$ degrees of freedom. Thus the function $\sqrt{n}(\bar{Y} - \mu)/s$ is pivotal for μ.

Hence, as in Section 20.2.1, we can construct a confidence interval at level $1 - \alpha$ for μ, which is symmetric about $\bar{Y}$ and whose boundaries are $\bar{Y} \pm ts\sqrt{n}$, where t is the $(1-\alpha/2)$-th quantile of the Student distribution with $n-1$ degress of freedom. Note that the length of the confidence interval thus obtained is $2ts/\sqrt{n}$ and is therefore random, unlike in Section 20.2.1.

20.2.3 Confidence Region for the Mean and the Variance of a Normal Distribution

We consider the same model as in Section 20.2.2 but we are now interested in the two-dimensional vector $\theta = (\mu, \sigma^2)'$. A possible pivotal

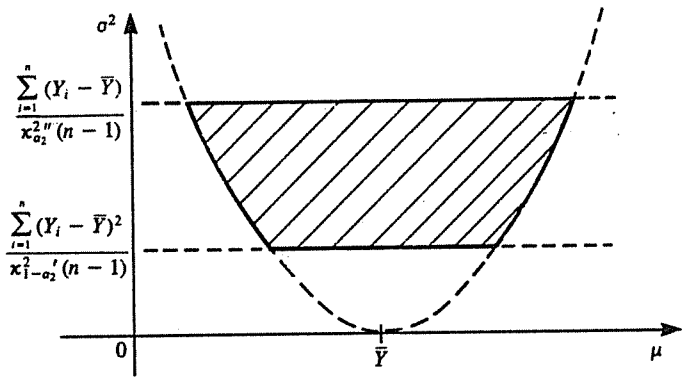


Figure 20.1: Confidence Region

function is the two-dimensional function whose components are

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \text{ and } \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2.$$

From Theorem B.7 we know that the joint probability distribution of this pair of functions is the product of the $N(0,1)$ distribution and the $\chi^2(n-1)$ distribution. This joint distribution is independent of θ as required.

Let α_1 and α_2 satisfy $\alpha = \alpha_1+\alpha_2-\alpha_1\alpha_2$. Thus we obtain a confidence region at level $1-\alpha$ for (μ,σ^2) of the form

$$\begin{cases} \dfrac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\chi^2_{1-\alpha_2'}(n-1)} \leq \sigma^2 \leq \dfrac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\chi^2_{\alpha_2''}(n-1)}, \\ \dfrac{n(\bar{Y}-\mu)^2}{\sigma^2} \leq u^2, \end{cases} \tag{20.1}$$

where α_2' and α_2'' are such that $\alpha_2 = \alpha_2' + \alpha_2''$, and u is the $(1-\alpha_1/2)$-th quantile of the $N(0,1)$ distribution. The boundaries of the region defined by the inequalities (20.1) are given by a parabola and two straight lines. This region is illustrated in Figure 20.1.

20.2.4 Confidence Region for the Parameters of a Linear Model

a) Confidence Interval

We consider the linear model

$$Y = \mathbf{X}b + u,$$

where u is a vector of random errors distributed as $N(0, \sigma^2\mathbf{I})$ and $\mathbf{X}$ is a constant $(n \times K)$ matrix of rank K.

We are interested in the i-th component b_i of b. Let $\hat{b}_i$ denote the i-th component of the ordinary least squares estimator $\hat{b}$ of b. We know that $\hat{b}_i$ follows a $N(b_i, \sigma^2 a^{ii})$ distribution, where a^{ii} is the i-th diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. Moreover, the statistic $(n-K)\hat{\sigma}^2 = \|Y - \mathbf{X}\hat{b}\|^2$ follows a $\chi^2(n-K)$ distribution and is independent of $\hat{b}_i$. Hence, $(\hat{b}_i - b_i)/\hat{\sigma}(a^{ii})^{1/2}$, which follows a Student distribution with $n-K$ degrees of freedom, is a pivotal function for b_i. It follows that a confidence interval at level $1-\alpha$ for b_i, which is symmetric about $\hat{b}_i$, is defined by the boundaries

$$\hat{b}_i \pm t\hat{\sigma}_i,$$

where t is the $(1-\alpha/2)$-th quantile of the Student distribution with $n-K$ degrees of freedom and $\hat{\sigma}_i = \hat{\sigma}(a^{ii})^{1/2}$.

b) Confidence Region

We consider the same model as above but we are now interested in the parameter vector b in the deterministic part of the linear model. It is easy to see that the quadratic form

$$\frac{1}{K\hat{\sigma}^2}(\hat{b}-b)'\mathbf{X}'\mathbf{X}(\hat{b}-b)$$

follows a Fisher–Snedecor distribution with K and $n-K$ degrees of freedom. Therefore this quadratic form is a pivotal function for b. Hence a confidence region at level $(1-\alpha)$ for b is given by the ellipsoid of $\mathbb{R}^K$ defined by

$$\frac{1}{K\hat{\sigma}^2}(\hat{b}-b)'\mathbf{X}'\mathbf{X}(\hat{b}-b) \leq c,$$

where c is the $(1-\alpha)$-th quantile of the Fisher–Snedecor distribution with K and $n-K$ degrees of freedom.

20.3 Conservative Confidence Regions

Definition 20.3: *A conservative confidence region at level* $1-\alpha$ *for* $g(\theta)$ *is a family of subsets* $\{S(y), y \in \mathcal{Y}\}$ *of* $G = g(\Theta)$ *such that*

$$\forall\, \theta \in \Theta,\ P_\theta(S(y) \ni g(\theta)) \geq 1-\alpha.$$

Conservative confidence regions are frequently used when:

(i) $g(\theta)$ is multidimensional with dimension q,

(ii) a confidence region $S_j(Y)$ at level $1-\alpha_j$ is available for each component $g_j(\theta)$, $j = 1, \ldots, q$, of $g(\theta)$, and

(iii) the events $\{S_j(\theta) \ni g_j(\theta)\}$ are not mutually independent.

Property 20.1: *Let* $S_j(Y)$ *be a confidence region for* $g_j(\theta)$ *at level* $1-\alpha_j$, $j = 1, \ldots, q$. *The region* $S(Y) = \prod_{j=1}^{q} S_j(Y)$ *is a conservative confidence region for* $g(\theta)$ *at level* $1 - \sum_{j=1}^{q} \alpha_j$, *where it is assumed that* $\sum_{j=1}^{q} \alpha_j \leq 1$.

PROOF: We have

$$\begin{aligned} P_\theta\left(\prod_{j=1}^{q} S_j(Y) \ni g(\theta)\right) &= P_\theta\left(\bigcap_{j=1}^{q} \{S_j(Y) \ni g_j(\theta)\}\right) \\ &= 1 - P_\theta\left(\bigcup_{j=1}^{q} \{S_j(Y) \not\ni g_j(\theta)\}\right) \\ &\geq 1 - \sum_{j=1}^{q} P_\theta(S_j(Y) \not\ni g_j(\theta)) \\ &= 1 - \sum_{j=1}^{q} \alpha_j. \end{aligned}$$

□

Example 20.1: Consider again the case of a normal distribution with unknown mean and variance (see Section 20.2.3). We know that a confidence interval for μ at level $1-\alpha_1$ is

$$\left[\bar{Y} - \frac{t_1 s}{\sqrt{n}}, \bar{Y} + \frac{t_1 s}{\sqrt{n}}\right],$$

where t_1 is the $(1-\alpha_1/2)$-th quantile of the Student distribution with $n-1$ degrees of freedom.

Similarly, we can construct a confidence interval for σ^2 at level $1-\alpha_2$ by using the pivotal function $1/\sigma^2 \sum_{i=1}^n (Y_i - \bar{Y})^2$, which follows a $\chi^2(n-1)$ distribution. Thus, we obtain a confidence interval of the form

$$\left[\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\chi^2_{1-\alpha_2'}(n-1)}, \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\chi^2_{\alpha_2''}(n-1)} \right],$$

where $\alpha_2' + \alpha_2'' = \alpha_2$.

The product of these two confidence intervals is a rectangle. From Property 20.1 this rectangle is a conservative confidence region for (μ, σ^2) at level $1-\alpha_1-\alpha_2$.

Example 20.2: Consider the linear model defined in Section 20.2.4-a). Consider two components of b, for instance, the first two components b_1 and b_2. For $i=1,2$, we know that a confidence interval for b_i at level $1-\alpha_i$ is

$$\left[\hat{b}_i - t_i \hat{\sigma} (a^{ii})^{1/2},\ \hat{b}_i + t_i \hat{\sigma} (a^{ii})^{1/2} \right], \quad i=1,2,$$

where t_i is the $(1-\alpha_i/2)$-th quantile of the Student distribution with $n-K$ degrees of freedom.

From Property 20.1 the rectangle obtained by taking the product of the preceding two intervals is a conservative confidence region for $(b_1, b_2)'$ at level $1-\alpha_1-\alpha_2$. For instance, for n sufficiently large and $\alpha_1 = \alpha_2 = 5\%$ we have $t_1 = t_2 \approx 2$. Let $\hat{\sigma}_i = \hat{\sigma}(a^{ii})^{1/2}$ denote the estimated standard error for the coefficient b_i. This conservative confidence rectangle for (b_1, b_2) at level 90% is

$$[\hat{b}_1 - 2\hat{\sigma}_1, \hat{b}_1 + 2\hat{\sigma}_1] \times [\hat{b}_2 - 2\hat{\sigma}_2, \hat{b}_2 + 2\hat{\sigma}_2].$$

20.4 Optimality of a Confidence Region

The preceding examples show that there can exist an infinite number of confidence regions at a given level. For instance, for the mean of a normal distribution with known variance (see Section 20.2.1), we can construct an infinite number of confidence intervals not necessarily symmetric about $\bar{Y}$. Specifically, we have

$$\forall \theta,\ P_\theta \left(u_1 \leq \sqrt{n} \frac{\bar{Y} - \theta}{\sigma_0} \leq u_2 \right) = 1 - \alpha,$$

for every segment $[u_1, u_2]$ with probability $1 - \alpha$ for the $N(0,1)$ distribution. Then the corresponding confidence intervals for θ at level $1 - \alpha$ are of the form

$$\left[\bar{Y} + \frac{u_1\sigma_0}{\sqrt{n}}, \bar{Y} + \frac{u_2\sigma_0}{\sqrt{n}}\right].$$

To choose among this infinite number of confidence intervals we may think of selecting the interval with the smallest length. It is easy to see that such a criterion leads to selecting the interval that is symmetric about $\bar{Y}$ (see Exercise 20.4). The problem with such a criterion, however, is that it is not sufficiently general. It is therefore necessary to consider other selection criteria.

To do so, we first define a preference relation on the set of confidence regions for $g(\theta)$ at a given level $1 - \alpha$. This relation will thus define a preorder on this set of confidence regions.

Definition 20.4: *Consider two confidence regions S and S' for $g(\theta)$ at level $1 - \alpha$. Then, S is said to be preferred to S' or more precise than S', which is denoted by $S \succeq S'$, if*

$$\forall\theta,\ \theta^*, P_\theta(S(Y) \ni g(\theta^*)) \leq P_\theta(S'(Y) \ni g(\theta^*)).$$

When $g(\theta) = g(\theta^*)$, the two sides of the preceding inequality are both equal to $1 - \alpha$. When $g(\theta) \neq g(\theta^*)$, then the preceding inequality says that, for every value of the parameter θ, the probability that S contains an erroneous value of g is smaller or equal to the analogous probability for S'.

The concept of preference defined above introduces a corresponding concept of optimality over confidence regions.

Definition 20.5: *A confidence region at level $1 - \alpha$ for $g(\theta)$ is said to be uniformly most precise (UMP) at level $1 - \alpha$ if it is more precise than any other confidence region at level $1 - \alpha$ for $g(\theta)$.*

The optimality criterion thus defined is actually very demanding. As a result, there does not exist a UMP region, in general. As in the theory of point estimation or in the theory of hypothesis tests, it is necessary to introduce other principles. Here, we shall restrict ourselves to the study of the most frequently used principle, namely, the principle of unbiased confidence regions.

Definition 20.6: *A confidence region S at level $1-\alpha$ for $g(\theta)$ is unbiased if*

$$\forall\theta,\ \theta^*, P_\theta(S(Y) \ni g(\theta)) \geq P_\theta(S(Y) \ni g(\theta^*)).$$

Note that the left-hand side of the inequality is equal to $1-\alpha$. Hence this inequality says that, for every θ, the probability that S contains the true value of g is larger or equal to the probability that S contains any erroneous value for g.

In the class of unbiased confidence regions at level $1-\alpha$ we can ask whether there exists an optimal region in the following sense.

Definition 20.7: *A confidence region at level $1-\alpha$ for $g(\theta)$ is uniformly most powerful and unbiased (UMPU) if it is unbiased and more precise than any other unbiased confidence region at level $1-\alpha$.*

Our problem now reduces to deriving some results that can be useful for constructing UMPU confidence regions. As a matter of fact, we shall see that the search for a UMPU region reduces to the search for an UMPU test. Hence we can apply the theory of unbiased tests presented in Chapter 16.

Let $\{S(y); y \in \mathcal{Y}\}$ be a confidence region for $g(\theta)$ at level $1-\alpha$. To this family of subsets of $G = g(\Theta)$ indexed by $y \in \mathcal{Y}$ we can associate a family of subsets of $\mathcal{Y}$ indexed by $d \in G$ defined by

$$A(d) = \{y : S(y) \in d\}.$$

Thus we have the following equivalence

$$y \in A(d) \Leftrightarrow d \in S(y). \tag{20.2}$$

Since $\{S(y); y \in \mathcal{Y}\}$ is a confidence region for $g(\theta)$ at level $1-\alpha$, we have

$$\forall \theta \in \Theta,\ P_\theta(S(y) \ni g(\theta)) = 1-\alpha.$$

Hence, $\forall d$ and $\forall \theta$ such that $g(\theta) = d$, we have

$$P_\theta(S(Y) \ni d) = 1-\alpha.$$

From (20.2) this implies that $\forall d$ and $\forall \theta$ such that $g(\theta) = d$ we have

$$P_\theta(Y \in A(d)) = 1-\alpha.$$

Let $H_0(d) = \{\theta : g(\theta) = d\}$. Then we obtain

$$\forall d, \forall \theta \in H_0(d) :\ P_\theta(Y \in A(d)) = 1-\alpha. \tag{20.3}$$

As a result, for every fixed d ($\in G$), the subset $A(d)$ can be interpreted as an acceptance region, i.e., the complement of a critical region of a test

at level α for the testing problem defined by the null hypothesis $H_0(d)$. Moreover, such a test is α-similar on the subset defined by $H_0(d)$ (see Definition 16.3).

Conversely, let $\{A(d), d \in G\}$ be a family of subsets of $\mathcal{Y}$ such that, for every $d \in G$, the subset $A(d)$ is an acceptance region of an α-similar test on $H_0(d)$ at level α of the null hypothesis $H_0(d)$. Then the family of subsets of $G = g(\Theta)$ defined by the equivalence (20.2) is a confidence region for $g(\theta)$ at level $1 - \alpha$ since

$$\forall \theta,\ P_\theta(S(Y) \ni g(\theta)) = P_\theta(Y \in A(g(\theta))) = 1 - \alpha.$$

The next property shows that the equivalence (20.2) closely links preference relations over confidence regions to preference relations over tests.

Property 20.2: *Let S and S' be two confidence regions for $g(\theta)$ at level $1 - \alpha$. Let A and A' be the corresponding families of subsets defined by (20.2). Then, the confidence region S is more precise than S' if and only if, $\forall d \in G = g(\Theta)$, the test defined by the acceptance region $A(d)$ is more powerful than the test defined by the acceptance region $A'(d)$ for testing the null hypothesis $H_0(d) = \{\theta : g(\theta) = d\}$.*

PROOF: The confidence region S is more precise than S' if and only if

$$\forall \theta, \theta^*,\ P_\theta(S(Y) \ni g(\theta^*)) \leq P_\theta(S'(Y) \ni g(\theta^*)),$$

i.e., if and only if

$$\forall \theta, \theta^*,\ P_\theta(Y \in A(g(\theta^*))) \leq P_\theta(Y \in A'(g(\theta^*))).$$

The latter inequality is satisfied with an equality for every pair (θ, θ^*) satisfying $\theta \in H_0(g(\theta^*))$ because the two sides are both equal to $1 - \alpha$ in this case. Thus this inequality is equivalent to

$$\forall d, \forall \theta \notin H_0(d),\ P_\theta(Y \in A(d)) \leq P_\theta(Y \in A'(d)),$$

i.e., to

$$\forall d, \forall \theta \notin H_0(d),\ P_\theta(Y \in {}^cA(d)) \geq P_\theta(Y \in {}^cA'(d)).$$

This means that, for every d, the test defined by the critical region ${}^cA(d)$ is more powerful than the test defined by the critical region ${}^cA'(d)$. □

Similarly, as the next property indicates, the concepts of unbiased confidence regions and unbiased tests are closely related through the equivalence (20.2).

Property 20.3: *Let S be a confidence region for $g(\theta)$ at level $1-\alpha$ and let A be the corresponding family of subsets defined by (20.2). Then, S is unbiased if and only if, $\forall d \in G$, the test defined by the acceptance region $A(d)$ is unbiased for testing the null hypothesis $H_0(d)$.*

PROOF: The confidence region S is unbiased if and only if

$$\forall \theta, \theta^*,\ P_\theta(S(Y) \ni g(\theta)) \geq P_\theta(S(Y) \ni g(\theta^*)),$$

i.e., if and only if

$$\forall \theta, \theta^*,\ 1-\alpha \geq P_\theta(S(Y) \ni g(\theta^*)).$$

In terms of critical regions, the latter inequality becomes

$$\forall d, \forall \theta \notin H_0(d) : P_\theta(Y \in^c A(d)) \geq \alpha.$$

As desired, this means that, for every d, the test defined by the critical region ${}^cA(d)$ is unbiased. □

From Properties 20.2 and 20.3 it follows that the search for a UMPU confidence region reduces to the search for a UMPU test that is α-similar on the subset defined by the null hypothesis.

20.5 Examples of UMPU Confidence Regions

20.5.1 Interval Estimation of the Mean of a Normal Distribution With Unknown Variance

We consider the sampling model $(\mathbb{R}, N(\mu, \sigma^2))^{\otimes n}$, where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$. We have $\theta = (\mu, \sigma^2)'$ and we seek a UMPU confidence region for $g(\theta) = \mu$ at level $1-\alpha$. The null hypothesis associated with this problem is $H_0(\mu_0) = \{(\mu, \sigma^2) : \mu = \mu_0\}$. The theory of UMPU tests with nuisance parameters presented in Section 16.1.3 implies that a UMPU test at level α for testing $H_0(\mu_0)$ is defined by the acceptance region

$$A(\mu_0) = \left\{ y : \frac{\sqrt{n}|\bar{y} - \mu_0|}{s} \leq t_{1-\alpha/2}(n-1) \right\}, \tag{20.4}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

and $t_{1-\alpha/2}(n-1)$ is the $(1-\alpha/2)$-th quantile of the Student distribution with $n-1$ degrees of freedom. Moreover, such a test is α-similar on the subset defined by $H_0(\mu_0)$. Therefore the equivalence (20.2) implies that a UMPU confidence region at level $1-\alpha$ for $g(\theta)$ is obtained by solving the inequality in (20.4) for μ_0. Specifically, we obtain

$$S(y) = \left\{\mu : \bar{y} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) \leq \mu \leq \bar{y} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right\}.$$

Hence the confidence region found in Section 20.2.2 on the basis of a pivotal function is UMPU.

20.5.2 Confidence Interval for the Parameters of a Linear Model

We consider the linear model $Y = \mathbf{X}b + u$, where u is a random disturbance distributed as $N(0, \sigma^2\mathbf{I})$ and $\mathbf{X}$ is a constant $(n \times K)$ matrix of rank K. We seek a UMPU confidence interval at level $1-\alpha$ for the i-th component b_i of b.

The corresponding null hypothesis is $H_0(b_i^0) = \{b : b_i = b_i^0\}$. In Section 16.1.3-d) we have shown that there exist a UMPU test for testing such a null hypothesis at level α and that this test is α-similar on the subset defined by $H_0(b_i^0)$. This test is defined by the acceptance region

$$A(b_i^0) = \left\{y : \frac{|\hat{b}_i - b_i^0|}{\hat{\sigma}_i} \leq t_{1-\alpha/2}(n-K)\right\}, \tag{20.5}$$

where $\hat{b}_i$ is the ordinary least squares estimator of b_i, $\hat{\sigma}_i = \hat{\sigma}(a^{ii})^{1/2}$ is the corresponding estimator of the standard error of $\hat{b}_i$, and $t_{1-\alpha/2}(n-K)$ is the $(1-\alpha/2)$-th quantile of the Student distribution with $n-K$ degrees of freedom.

Then the equivalence (20.2) and the equality (20.5) imply that a UMPU confidence interval at level $1-\alpha$ for b_i is

$$S(y) = \left\{b_i : \hat{b}_i - \hat{\sigma}_i t_{1-\alpha/2}(n-K) \leq b_i \leq \hat{b}_i + \hat{\sigma}_i t_{1-\alpha/2}(n-K)\right\}.$$

20.6 Asymptotic Confidence Regions

In the chapters dealing with point estimation we have frequently considered the asymptotic properties of point estimators $\hat{\theta}_n$ of the parameter

vector θ. In particular, we have established various asymptotic normality properties of the kind

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma(\theta)), \tag{20.6}$$

where $\Sigma(\theta)$ is a $p \times p$ variance covariance matrix. Such a kind of results can be used to construct asymptotic confidence regions for $g(\theta)$. Specifically, this can be done by introducing the concept of asymptotic pivotal functions defined next.

Definition 20.8: *Let $Y^{(n)} = (Y_1, \ldots, Y_n)$ be the identity statistic of a statistical model associated with n observations. An asymptotic pivotal function for $g(\theta)$ is a sequence of functions ν_n such that the asymptotic distribution of $\nu_n(Y^{(n)}, g(\theta))$ is independent of θ as n increases to infinity.*

Let Π_ν denote the asymptotic distribution of $\nu_n(Y^{(n)}, g(\theta))$. For n sufficiently large, we can approximate the distribution of $\nu_n(Y^{(n)}, g(\theta))$ by Π_ν. Thus, we can construct a confidence region for $g(\theta)$ as in Section 20.1 where a pivotal function was available.

For instance, consider the asymptotic normality situation (20.6), which arises frequently in practice. Suppose that we are interested in the parameter θ itself so that $g(\theta) = \theta$. Then, if $\Sigma(\theta)$ is nonsingular and continuous in θ, we can consider the function

$$\nu_n = n(\hat{\theta}_n - \theta)' \Sigma^{-1}(\hat{\theta}_n)(\hat{\theta}_n - \theta)$$

as a pivotal function for θ. This quantity is asymptotically distributed as a χ^2 distribution with p degrees of freedom. Such an asymptotic pivotal function allows us to construct a confidence ellipsoid at asymptotic level $1 - \alpha$ defined by

$$\{\theta : n(\theta - \hat{\theta}_n)' \Sigma^{-1}(\hat{\theta}_n)(\theta - \hat{\theta}_n) \leq \chi^2_{1-\alpha}(p)\}, \tag{20.7}$$

where $\chi^2_{1-\alpha}(p)$ is the $(1 - \alpha)$-th quantile of the χ^2 distribution with p degrees of freedom. Such a confidence ellipsoid is centered around the estimator $\hat{\theta}_n$. See Figure 20.2 for the case where the dimension p of θ is two.

Suppose now that we are interested in a subvector of dimension q of θ. Then we obtain a result analogous to (20.7), where θ and $\hat{\theta}_n$ are replaced by their corresponding subvectors, $\Sigma^{-1}(\hat{\theta}_n)$ is replaced by the inverse of the corresponding block of $\Sigma(\hat{\theta}_n)$, and p is replaced by q.

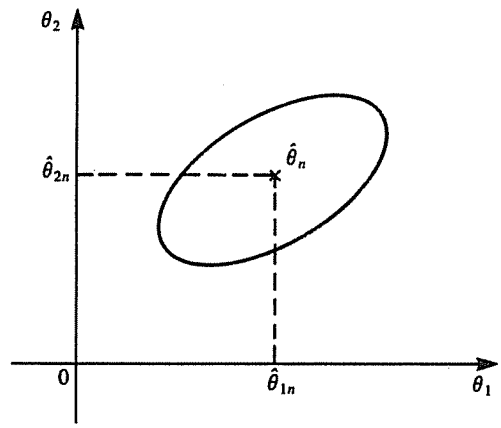


Figure 20.2: Asymptotic Confidence Region

More generally, suppose that we are interested in a function $g(\theta)$ of the parameter vector θ, where g is continuously differentiable on $\mathbb{R}^p$ with values on $\mathbb{R}^q$. From Property B.69 we know that the asymptotic normality of $\hat{\theta}_n$ implies that

$$\sqrt{n}(g(\hat{\theta}_n - g(\theta)) \xrightarrow{d} N(0, \mathbf{\Delta}(\theta)\mathbf{\Sigma}(\theta)\mathbf{\Delta}'(\theta)),$$

where $\mathbf{\Delta}(\theta)$ is the $(q \times p)$ matrix of partial derivatives $\partial g_i(\theta)/\partial\theta_j$, $i = 1, \ldots, q$, $j = 1, \ldots, p$. Consequently, if $\mathbf{\Delta}(\theta)$ is of rank q, then we have

$$n(g(\hat{\theta}_n) - g(\theta))'(\mathbf{\Delta}(\theta)\mathbf{\Sigma}(\theta)\mathbf{\Delta}'(\theta))^{-1}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \chi^2(q) \quad (20.8)$$

and

$$n(g(\hat{\theta}_n) - g(\theta))'(\mathbf{\Delta}(\hat{\theta}_n)\mathbf{\Sigma}(\hat{\theta}_n)\mathbf{\Delta}'(\hat{\theta}_n))^{-1}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \chi^2(q). \quad (20.9)$$

It follows that a confidence ellipsoid for $g(\theta)$ at asymptotic level $1 - \alpha$ is given by the set of values $g(\theta)$ satisfying

$$n(g(\theta) - g(\hat{\theta}_n))'(\mathbf{\Delta}(\hat{\theta}_n)\mathbf{\Sigma}(\hat{\theta}_n)\mathbf{\Delta}'(\hat{\theta}_n))^{-1}(g(\theta) - g(\hat{\theta}_n)) \leq \chi^2_{1-\alpha}(q). \quad (20.10)$$

20.7 Examples of Asymptotic Confidence Regions

The results presented in Section 20.6 are quite general. For instance, they apply when $\hat{\theta}_n$ is obtained by maximizing an arbitrary statistical objective function $L_n(\theta)$. This is the case for the maximum likelihood estimator (see Chapter 7), for M-estimators and in particular pseudo maximum likelihood estimators and quasi generalized pseudo maximum likelihood estimators (see Chapter 8), and for moment estimators and their generalizations such as asymptotic least squares estimators and generalized methods of moments estimators (see Chapter 9). In all those cases we have seen that, under suitable regularity conditions and normalization of $L_n(\theta)$, we have

$$\frac{1}{\sqrt{n}}\frac{\partial L_n}{\partial \theta} \xrightarrow{d} N(0, \mathcal{I}(\theta))$$

and

$$-\frac{1}{n}\frac{\partial^2 L_n}{\partial \theta \partial \theta'} \xrightarrow{a.s.} \mathcal{J}(\theta).$$

These properties imply that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \mathcal{J}^{-1}(\theta)\mathcal{I}(\theta)\mathcal{J}^{-1}(\theta)).$$

Thus the conditions of the preceding section are satisfied with

$$\mathcal{J}^{-1}(\theta)\mathcal{I}(\theta)\mathcal{J}^{-1}(\theta) = \mathbf{\Sigma}(\theta).$$

Futhermore, in Chapter 7, 8, and 9 we have seen that the equality $\mathcal{I}(\theta) = \mathcal{J}(\theta)$ holds in many important cases. These cases include maximum likelihood, quasi generalized pseudo maximum likelihood, optimal asymptotic least squares, and optimal generalized methods of moments. When such an equality holds, then $\mathbf{\Sigma}(\theta)$ is equal to $\mathcal{J}^{-1}(\theta)$. Then a consistent estimator of $\mathbf{\Sigma}(\theta)$ is

$$\left(-\frac{1}{n}\frac{\partial^2 L_n(\hat{\theta}_n)}{\partial \theta \partial \theta'}\right)^{-1}. \tag{20.11}$$

In particular, it follows that an asymptotic confidence region for θ is

$$\left\{\theta : -(\theta - \hat{\theta}_n)'\frac{\partial^2 L_n(\hat{\theta}_n)}{\partial \theta \partial \theta'}(\theta - \hat{\theta}_n) \leq \chi^2_{1-\alpha}(p)\right\}.$$

This asymptotic confidence region is easy to construct since the Hessian $\partial^2 L_n(\hat{\theta}_n)/\partial\theta\partial\theta'$ is, in general, a by-product of commonly used numerical maximization algorithms of $L_n(\theta)$. In addition, this confidence ellipsoid can be interpreted as the set of θ for which the quadratic approximation of L_n around $\hat{\theta}_n$ is larger than

$$L_n(\hat{\theta}_n) - \frac{1}{2}\chi^2_{1-\alpha}(p).$$

This follows from the equality

$$\left\{\theta : -(\theta - \hat{\theta}_n)' \frac{\partial^2 L_n(\hat{\theta}_n)}{\partial\theta\partial\theta'}(\theta - \hat{\theta}_n) \leq \chi^2_{1-\alpha}(p)\right\} =$$

$$\left\{\theta : L_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)' \frac{\partial^2 L_n(\hat{\theta}_n)}{\partial\theta\partial\theta'}(\theta - \hat{\theta}_n) \geq L_n(\hat{\theta}_n) - \frac{1}{2}\chi^2_{1-\alpha}(p)\right\}.$$

20.8 Bayesian Confidence Regions

Within the Bayesian approach (see Chapters 4 and 12) the parameter space Θ is endowed with a prior distribution Π. Then the observation y modifies this prior distribution into a posterior distribution $\Pi^{\theta|y}$, which is the conditional distribution of θ given $Y = y$.

Let $\Pi^{g(\theta)|y}$ denote the conditional distribution of the function $g(\theta)$ of the parameter. This conditional distribution is the image of $\Pi^{\theta|y}$ by g.

Definition 20.9: *A Bayesian confidence region at level* $1-\alpha$ *for* $g(\theta)$ *is a family of the subset of* $G = g(\theta)$ *indexed by* $y \in \mathcal{Y}$, *denoted* $\{S(y), y \in \mathcal{Y}\}$, *such that*

$$\forall y \in \mathcal{Y},\ \Pi^{g(\theta)|y}(S(y)) = 1 - \alpha.$$

Confidence regions have quite different interpretations whether one adopts the classical framework (Definition 20.1) or the Bayesian framework (Definition 20.9). Specifically, within the classical framework we have

$$P_\theta(S(Y) \ni g(\theta)) = 1 - \alpha, \tag{20.12}$$

while within the Bayesian framework we have

$$\Pi^{\theta|y}(S(y) \ni g(\theta)) = 1 - \alpha. \tag{20.13}$$

An important difference is that $S(Y)$ is random while $g(\theta)$ is fixed in (20.12). In contrast, $S(y)$ is fixed while $g(\theta)$ is random in (20.13).

For instance, when $g(\theta)$ is a scalar, a given interval $S(y)$ such as $[1.5; 1.8]$ will have two entirely different interpretations whether one adopts the classical framework or the Bayesian framework. Within the former framework the interval $[1.5; 1.8]$ is the realization of a random interval, that contains with probability $1-\alpha$ the true value $g(\theta)$ *ex ante*, i.e., prior to the random experiment. On the other hand, within the latter framework the interval $[1.5; 1.8]$ is an interval that contains with probability $1-\alpha$ the value $g(\theta)$ *ex post*, i.e., after the random experiment is realized.

To illustrate Definition 20.8, we consider the following example.

Example 20.3: Let $Y = (Y_1, \ldots, Y_n)$ be a random sample drawn from the normal distribution $N(\theta, \sigma_0^2)$, where σ_0^2 is known. We consider a diffuse prior distribution defined by $\pi(\theta) \propto cst$, $\theta \in I\!R$. The likelihood function is

$$\begin{aligned} l(y;\theta) &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma_0^2}(\bar{y}-\theta)^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\}. \end{aligned}$$

Hence the posterior distribution corresponding to the above diffuse prior distribution is proportional to

$$\exp\left\{-\frac{n}{2\sigma_0^2}(\theta-\bar{y})^2\right\}.$$

That is, the posterior distribution is the $N(\bar{y}, \sigma_0^2/n)$ distribution. Thus, a Bayesian confidence interval at level $1-\alpha$ for θ symmetric about $\bar{y}$ is defined by the boundaries $\bar{y} \pm u\sigma_0/\sqrt{n}$, where u is the $(1-\alpha/2)$-th quantile of the standard normal distribution $N(0,1)$. Note that we obtain again the confidence interval found in Section 20.2.1.

20.9 Exercises

EXERCISE 20.1: Consider the sampling model $(I\!R, N(\mu, \sigma^2))^{\otimes n}$, where $\mu \in I\!R$ and $\sigma^2 \in I\!R^+$ are unknown. Consider the pair (μ, σ^2). Show that $\sum_{i=1}^{n}(Y_i-\mu)^2/\sigma^2$ is a pivotal function. Determine the shape of the confidence regions associated with this pivotal function.

EXERCISE 20.2: Generalize the results of the example in Section 20.2.4-c) to the case where one is interested in a subvector of b or in a vector of linear forms in b.

EXERCISE 20.3: Consider the linear model studied in Section 20.2.4. Propose a confidence interval for σ^2.

EXERCISE 20.4: Consider a centered normal distribution defined on the real line. For a given probability content, show that the interval of minimal length is an interval centered around zero.

EXERCISE 20.5: Let $l(y)$ be a probability density with respect to the Lebesgue measure on $\mathbb{R}$. Suppose that l is continuous. For a given probability content, show that the interval $[a, b]$ of minimal length is such that $l(a) = l(b)$.

EXERCISE 20.6: Consider the sampling model $(\{0,1\}^n, B(1,\theta)^{\otimes n})$, where $B(1,\theta)$ is the Bernoulli distribution defined by $P_\theta(\{1\}) = \theta$ and $P_\theta(\{0\}) = 1 - \theta$. Show that

$$\frac{\sqrt{n}(\bar{Y} - \theta)}{\sqrt{\theta(1-\theta)}}$$

is an asymptotic pivotal function. Then find an asymptotic confidence interval at asymptotic level $1 - \alpha$ for θ.

Answer the same questions using the function

$$\frac{\sqrt{n}(\bar{Y} - \theta)}{(\bar{Y}(1-\bar{Y}))^{1/2}}.$$

EXERCISE 20.7: Let $\hat{\theta}_n$ be a consistent and $\sqrt{n}$-asymptotically normal estimator of the bidimensional parameter vector $\theta' = (\theta_1, \theta_2)$. Let Σ denote the asymptotic variance covariance matrix of $\sqrt{n}(\hat{\theta}_n - \theta)$. For $n = 100$,

$$\hat{\theta}_n = \begin{pmatrix} 10 \\ 8 \end{pmatrix}, \Sigma = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix},$$

find an asymptotic confidence region for $(\theta_1 - \theta_2, \theta_1^2 + \theta_2^2)$ at asymptotic level 0.95.

20.10 References

Cox, D.R. and Hinckley D.V. (1974). *Theoretical Statistics*, Chapman and Hall.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, Wiley.

Silvey, S.D. (1970). *Statistical Inference*, Chapman and Hall.

CHAPTER 21

Inequality Constraints: Estimation and Testing

In many estimation and testing situations the parameter vector of interest is subject to some implicit constraints of the form

$$g(\theta) \geq 0,$$

where g is a function from $\mathbb{R}^p$ to $\mathbb{R}^r$ with rank r and where the symbol "$g \geq a$" means that each component g_j of g is greater or equal to a with $a \in \mathbb{R}$. It is therefore necessary to develop some estimation methods and hypothesis tests that can take into account these inequality restrictions explicitly. The study and the application of such procedures are, however, more difficult than in the case of equality constraints.

In this chapter our goal is to give a simplified presentation of the main results in this area.

21.1 Some Examples

Inequality constraints arise typically in the following four situations:

(i) When a function, which is parameterized by θ, is constrained to be increasing (or decreasing) so that the derivatives of the function must be positive (or negative).

(ii) When a function of the parameter vector is interpreted as a probability (or as a vector of barycentric coordinates) so that the values of the function must be between 0 and 1.

(iii) When symmetric matrices parameterized by θ are supposed to be positive (or negative) semidefinite.

(iv) When conditions on the parameters of a dynamic model must be imposed to ensure the stability of the model.

Each of these situations are now considered in turn.

21.1.1 Monotony Constraints

Example 21.1: Consider a consumption model of the form

$$\log C_{lt} = a_l \log R_t + b_l + u_{lt}, \quad t = 1, \ldots, T,$$

where C_{lt} denotes the consumption expenditure on commodity l. Then imposing that consumption is increasing in income, R_t is equivalent to imposing the constraint $a_l \geq 0$. Moreover, imposing that the commodity is a luxury good, i.e., that the proportion of consumption to income C_{lt}/R_t is increasing in income, is equivalent to imposing that $a_l \geq 1$.

Example 21.2: Consider a macroeconomic model that contains an adjustment price equation of the form

$$P_t - P_{t-1} = \lambda(D_{t-1} - S_{t-1}),$$

where D and S denote the demand and the supply respectively. Then it is natural to impose that the change in price is an increasing function of the excess demand $D - S$, i.e., to impose the constraint $\lambda \geq 0$.

21.1.2 Barycentric Interpretation of Some Parameters

Example 21.3: Consider two statistical models $\mathcal{M}_1 = \{f_1(y), f_1 \in F_1\}$ and $\mathcal{M}_2 = \{f_2(y), f_2 \in F_2\}$ dominated by a measure μ. In some situations such as in Chapter 22 it may be interesting to construct a model that nests both $\mathcal{M}_1$ and $\mathcal{M}_2$. A general method for constructing such a nesting model is to consider the family of mixtures of the probability densities f_1 and f_2. The nesting model thus obtained is the collection of densities of the form

$$f(y; \theta, \lambda) = \lambda f_1(y; \theta) + (1 - \lambda) f_2(y; \theta).$$

The parameter λ must be between 0 and 1 to ensure that $f(y; \theta, \lambda)$ is a well-defined density.

Example 21.4: Consider the following classical disequilibrium model

$$\begin{cases} D_t &= ap_t + x_t b + u_t, \\ S_t &= \alpha p_t + z_t \beta + v_t, \\ Q_t &= \min(D_t, S_t), \end{cases}$$

with $a \leq 0$ and $\alpha \geq 0$. Such a model can be completed by introducing a price adjustment equation of the form

$$p_t - p_{t-1} = \lambda(D_t - S_t),$$

where $\lambda \geq 0$.

Define the equilibrium price p_t^e as a solution of

$$ap_t^e + x_t b + u_t = \alpha p_t^e + z_t \beta + v_t.$$

Then the price adjustment equation can be rewritten as

$$p_t - p_{t-1} = \lambda(a - \alpha)(p_t - p_t^e),$$

i.e.

$$p_t = \frac{1}{1 + \lambda(\alpha - a)} p_{t-1} + \frac{\lambda(\alpha - a)}{1 + \lambda(\alpha - a)} p_t^e.$$

Thus the current price p_t is a weighted average of the lagged price p_{t-1} and the equilibrium price p_t^e. It is convenient to introduce the new parameter $\gamma = 1/(1 + \lambda(\alpha - a))$. Then, because of the constraints $\alpha \geq 0$, $a \leq 0$, and $\lambda \geq 0$, the parameter γ must be between 0 and 1 as it can be readily verified.

21.1.3 Positive Definiteness of Symmetric Matrices

Example 21.5: The elements of a variance covariance matrix Σ (say) are frequently included in the list of parameters of a model. These elements are constrained by the conditions that Σ is symmetric and positive semidefinite.

For instance, if we consider the (3×3) case, then

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}.$$

Using the principal minors, the condition that this matrix is positive semidefinite, i.e., $\Sigma \succeq 0$, is equivalent to

$$\sigma_1^2 \geq 0, \ \sigma_1^2 \sigma_2^2 - \sigma_{12}^2 \geq 0, \ \det \Sigma \geq 0.$$

Example 21.6: Constraints similar to those in Example 21.5 arise in many demand models. Specifically, consider a consumer who chooses his levels of consumption for various commodities $l = 1, \ldots, L$ by maximizing his utility subject to his budget constraint. It can be shown that the demand functions

$$C_l = C_l(p, R; \theta), \quad l = 1, \ldots, L,$$

where p is the price vector and R is income, are not arbitrary. In particular, they must satisfy the budget constraint

$$\sum_{l=1}^{L} p_l C_l(p, R; \theta) = R.$$

Moreover, because they solve a constrained optimization problem, the demand equations must satisfy some conditions called the *Slutzky conditions* on the substitution matrix $\mathbf{S}$. The latter matrix is of size $(L \times L)$ whose generic term is

$$S_{kl} = \frac{\partial C_l(p, R; \theta)}{\partial p_k} + \frac{\partial C_l(p, R; \theta)}{\partial R} C_k(p, R; \theta).$$

Such a matrix must be symmetric and negative semidefinite. Such conditions clearly imply some inequality constraints on the parameter vector θ.

21.1.4 Stability Conditions

Example 21.7: Consider the following dynamic model

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \mathbf{A} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + X_t b + u_t.$$

It is interesting to impose some conditions on the elements of the matrix $\mathbf{A}$ so as to ensure that $y_t = (y_{1t}, y_{2t})'$ converges to some limit, when the exogenous variable X_t remains constant and there is no shock ($u_t \equiv 0$). Such a convergence arises if and only if $\mathbf{A}^t$ converges to the zero matrix as t increases to infinity. Such a condition is equivalent to the condition that the eigenvalues of the matrix $\mathbf{A}$ are all strictly smaller than 1.

21.2 Estimation Under Inequality Constraints

To illustrate the problem of estimation under inequality constraints, we first study the linear model, where only one parameter of the deterministic part of the model is constrained by an inequality.

21.2.1 Constrained Estimation of the Linear Model

We consider the Gaussian linear model

$$Y = \mathbf{X}\theta + u = X_1\theta_1 + \mathbf{X}_2\theta_2 + u,$$

where the error u follows the normal distribution $N(0, \sigma^2\mathbf{I})$ and where the first component of θ is constrained to be nonnegative, i.e., $\theta_1 \geq 1$.

a) General Form of the Constrained Estimator

The constrained ordinary least squares estimator of θ is defined as the solution to

$$\begin{cases} \min_\theta \|Y - X_1\theta_1 - \mathbf{X}_2\theta_2\|^2 \\ \text{subject to } \theta_1 \geq 0. \end{cases} \tag{21.1}$$

Let $\tilde{\theta}_n^c$ denote such an estimator. Let $\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{2n})'$ denote the (unconstrained) ordinary least squares (OLS) estimator of θ. Let $\hat{\theta}_n^0$ denote the OLS estimator of θ constrained by the equality $\theta_1 = 0$. We have

$$\hat{\theta}_n = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y,$$

$$\hat{\theta}_n^0 = \begin{pmatrix} 0 \\ \hat{\theta}_{2n}^0 \end{pmatrix} = \begin{pmatrix} 0 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y \end{pmatrix}.$$

The solution to the problem (21.1) has two distinct forms according to whether the inequality constraint $\theta_1 \geq 0$ is binding. Specifically, we obtain

$$\tilde{\theta}_n^c = \begin{cases} \hat{\theta}_n, \text{ if } \hat{\theta}_{1n} \geq 0, \\ \hat{\theta}_n^0, \text{ if } \hat{\theta}_{1n} \leq 0. \end{cases} \tag{21.2}$$

Thus, when the inequality constraint $\theta_1 \geq 0$ is not binding, the inequality constrained estimator coincides with the unconstrained OLS estimator

and, hence, follows a continuous distribution on $\mathbb{R}^p$. On the other hand, when the inequality constraint is binding, the inequality constrained estimator coincides with the equality constrained estimator and, hence, follows a continuous distribution on the subspace of dimension $p-1$ defined by $\theta_1 = 0$.

b) Relation Between the Inequality Constrained Estimator and the Unconstrained Estimator

In Chapter 10 we saw that the equality constrained estimator $\hat{\theta}_n^0$ depends on Y only through the unconstrained estimator $\hat{\theta}_n$. From equation (21.2) it follows that inequality constrained estimator $\tilde{\theta}_n^c$ possesses the same property. As a matter of fact, this can be shown directly using the optimization problem (21.1).

Denote by $\mathbf{M}$ the matrix associated with the projection on to the subspace generated by the columns of $\mathbf{X}$. We have

$$\begin{aligned} &\|Y - X_1\theta_1 - \mathbf{X}_2\theta_2\|^2 \\ &= \|\mathbf{M}(Y - X_1\theta_1 - \mathbf{X}_2\theta_2)\|^2 + \|(\mathbf{I}-\mathbf{M})(Y - X_1\theta_1 - \mathbf{X}_2\theta_2)\|^2 \\ &= \|\mathbf{X}(\hat{\theta}_n - \theta)\|^2 + \|(\mathbf{I}-\mathbf{M})Y\|^2. \end{aligned}$$

Thus the optimization problem (21.1) is equivalent to the problem

$$\begin{cases} \min_\theta\ (\hat{\theta}_n - \theta)'\mathbf{X}'\mathbf{X}(\hat{\theta}_n - \theta) \\ \text{subject to } \theta_1 \geq 0. \end{cases} \tag{21.3}$$

That is, we can replace the original linear model by the model

$$\hat{\theta}_n = \theta + \nu,\ E\nu = 0,\ V\nu = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Working with the latter problem, we obtain the following expressions for the inequality constrained estimators $\tilde{\theta}_{1n}^c$ and $\tilde{\theta}_{2n}^c$

$$\begin{pmatrix} \tilde{\theta}_{1n}^c \\ \tilde{\theta}_{2n}^c \end{pmatrix} = \begin{cases} \begin{pmatrix} \hat{\theta}_{1n} \\ \hat{\theta}_{2n} \end{pmatrix}, & \text{if } \hat{\theta}_{1n} \geq 0, \\ \begin{pmatrix} 0 \\ \hat{\theta}_{2n} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\hat{\theta}_{1n} \end{pmatrix}, & \text{if } \hat{\theta}_{1n} \leq 0. \end{cases} \tag{21.4}$$

See Exercise 10.1.

Remark 21.1: The inequality constraint on θ_1 can have an effect on the estimation of θ_2. Such an effect disappears when X_1 is orthogonal to the columns of $\mathbf{X}_2$.

Remark 21.2: The inequality constrained estimator of θ_2 can be written as

$$\tilde{\theta}_{2n}^c = \hat{\theta}_{2n} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1(\hat{\theta}_{1n} - \tilde{\theta}_{1n}^c),$$

i.e.

$$\tilde{\theta}_{2n}^c + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\tilde{\theta}_{1n}^c = \hat{\theta}_{2n} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\hat{\theta}_{1n}.$$

This equation shows that the constrained and unconstrained estimators of the parameter

$$\beta_2 = \theta_2 + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\theta_1$$

coincide. Such a remark is analogous to Remark 21.1 since β_2 can be interpreted as a parameter in the original linear model after orthogonalization of its explanatory variables.

c) Kühn and Tücker Multipliers

In the case of equality constraints the use of Lagrange multipliers was important for the construction of tests statistics. Similarly, in the case of inequality constraints it is possible to introduce corresponding multipliers. These are called *Kühn and Tücker multipliers.*

We consider again the above linear model. Let $\tilde{\lambda}_n$ denote the multiplier associated with the constraint $\theta_1 \geq 0$. The Lagrangian function is

$$\mathcal{L} = (Y - \mathbf{X}\theta)'(Y - \mathbf{X}\theta) - \lambda\theta_1.$$

The first-order conditions corresponding to the minimization problem under the inequality constraint $\theta_1 \geq 0$ includes some sign constraints and some exclusion restrictions. Namely, we have

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \theta_1} = 0 \Leftrightarrow -2X_1'(Y - X_1\tilde{\theta}_{1n}^c - \mathbf{X}_2\tilde{\theta}_{2n}^c) - \tilde{\lambda}_n = 0, \\ \dfrac{\partial \mathcal{L}}{\partial \theta_2} = 0 \Leftrightarrow -2\mathbf{X}_2'(Y - X_1\tilde{\theta}_{1n}^c - \mathbf{X}_2\tilde{\theta}_{2n}^c) = 0, \\ \text{sign constraints: } \tilde{\theta}_{1n}^c \geq 0, \tilde{\lambda}_n \geq 0, \\ \text{exclusion constraint: } \tilde{\lambda}_n\tilde{\theta}_{1n}^c = 0. \end{cases} \tag{21.5}$$

To solve this system we distinguish two cases according to whether $\tilde{\lambda}_n$ and $\tilde{\theta}_{1n}^c$ are zero or nonzero.

(i) Suppose that $\tilde{\lambda}_n > 0$. We necessarily have $\tilde{\theta}^c_{1n} = 0$ because of the exclusion constraint. Then, from the first two equations, we obtain

$$\begin{aligned} \tilde{\lambda}_n &= -2X_1'(Y - \mathbf{X}_2\tilde{\theta}^c_{2n}) \\ &= -2X_1'(Y - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y). \end{aligned}$$

(ii) Suppose that $\tilde{\theta}^c_{1n} > 0$. We necessarily have $\tilde{\lambda}_n = 0$. In this case it is easy to verify that $\tilde{\theta}^c_n = \hat{\theta}_n$. Hence $\tilde{\theta}^c_{1n} > 0$ if and only if $\hat{\theta}_{1n} > 0$.

(iii) A third case may arise. This is when $\tilde{\lambda}_n = 0$ and $\tilde{\theta}^c_{1n} = 0$ simultaneously. This implies

$$\begin{cases} X_1'(Y - \mathbf{X}_2\tilde{\theta}^c_{2n}) &= 0, \\ \mathbf{X}_2'(Y - \mathbf{X}_2\tilde{\theta}^c_{2n}) &= 0. \end{cases}$$

Eliminating $\tilde{\theta}^c_{2n}$, we obtain

$$X_1'(Y - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y) = 0.$$

Such a condition, however, is satisfied with zero probability since Y is continuously distributed and X_1 does not belong to the space generated by the columns of $\mathbf{X}_2$.

The linear model example shows that the system (21.5) determines simultaneously the value of the constrained estimator and the value of the multiplier. The multiplier can take two forms according to whether the inequality constraint $\theta_1 \geq 0$ is binding. Namely

$$\tilde{\lambda}_n = \begin{cases} -2X_1'(Y - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y), & \text{if } \hat{\theta}_{1n} \leq 0, \\ 0, & \text{if } \hat{\theta}_{1n} > 0. \end{cases}$$

Lastly, it remains to relate the Kühn and Tücker multiplier $\tilde{\lambda}_n$ to the Lagrange multiplier associated with the constraint $\theta_1 = 0$ and the score vector. The Lagrange multiplier associated with the optimization problem subject to the constraint $\theta_1 = 0$ is

$$\hat{\lambda}_n = -2X_1'(Y - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y).$$

Since

$$\hat{\theta}_{1n} = (X_1'(\mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2')X_1)^{-1}X_1'(\mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2')Y,$$

we see at once that $\hat{\lambda}_n \geq 0$ if and only if $\hat{\theta}_{1n} \leq 0$. This shows that the Kühn and Tücker multiplier $\tilde{\lambda}_n$ can be expressed directly as a function

of the Lagrange multiplier $\hat{\lambda}_n$ of the optimization problem subject to the equality constraint. Namely, we have

$$\tilde{\lambda}_n = \begin{cases} \hat{\lambda}_n, & \text{if } \hat{\lambda}_n \geq 0, \\ 0, & \text{if } \hat{\lambda}_n \leq 0. \end{cases} \tag{21.6}$$

Consider now the score $\partial L(Y; \tilde{\theta}_{1n}^c, \tilde{\theta}_{2n}^c)/\partial\theta_1$, where L denotes the objective function $L(Y; \theta_1, \theta_2) = \|Y - X_1\theta_1 - \mathbf{X}_2\theta_2\|^2$. We see at once that

$$\frac{\partial L(Y; \tilde{\theta}_{1n}^c, \tilde{\theta}_{2n}^c)}{\partial\theta_1} = \begin{cases} \dfrac{\partial L(Y; \hat{\theta}_{1n}, \hat{\theta}_{2n})}{\partial\theta_1} = 0, & \text{if } \hat{\theta}_{1n} \geq 0, \\ \dfrac{\partial L(Y; 0, \hat{\theta}_{2n})}{\partial\theta_1}, & \text{if } \hat{\theta}_{1n} \leq 0. \end{cases}$$

The latter form is nothing else other than that of $\tilde{\lambda}_n$. Thus

$$\tilde{\lambda}_n = \frac{\partial L(Y; \tilde{\theta}_n^c)}{\partial\theta_1}. \tag{21.7}$$

That is, the Kühn and Tücker multiplier $\tilde{\lambda}_n$ is equal to the "score" corresponding to the component θ_1 evaluated at the inequality constrained estimator $\tilde{\theta}_n^c$.

d) Finite Sample Distributions

The probability distributions of $\tilde{\theta}_{1n}^c$, $\tilde{\theta}_{2n}^c$, and $\tilde{\lambda}_n$ are obtained directly from their expressions in terms of the unconstrained estimators or scores. The probability distribution of $\tilde{\theta}_{1n}^c$ has a continuous part and a discrete part, which arises when $\tilde{\theta}_{1n}^c = 0$. The same remark applies to the probability distribution of $\tilde{\lambda}_n$ when $\tilde{\lambda}_n = 0$. On the other hand, the probability distribution of $\tilde{\theta}_{2n}^c$ is continuous.

(i) First, we consider the probability distribution of $\tilde{\theta}_{1n}^c$. We have

$$\tilde{\theta}_{1n}^c = \hat{\theta}_{1n} \mathbb{1}_{\hat{\theta}_{1n} \geq 0}.$$

Because the probability distribution of $\hat{\theta}_{1n}$ is the normal distribution $N(\theta_1^0, \sigma^2\omega_{11})$, it follows that the probability distribution of $\tilde{\theta}_{1n}^c$ is a truncated normal distribution. Specifically, we have

$$P(\tilde{\theta}_{1n}^c = 0) = P(\hat{\theta}_{1n} < 0) = \Phi\left(\frac{-\theta_1^0}{\sigma\sqrt{\omega_{11}}}\right),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. On the other hand, the probability distribution of $\tilde{\theta}_{1n}^c$ on $\mathbb{R}^{+*}$ has a continuous component corresponding to the probability distribution of $\hat{\theta}_{1n}$. Its density at $\tilde{\theta}_1$ is

$$\frac{1}{\sigma\sqrt{\omega_{11}}}\phi\left(\frac{\tilde{\theta}_1 - \theta_1^0}{\sigma\sqrt{\omega_{11}}}\right),$$

where ϕ denotes the probability density of Φ. Thus, the probability distribution of $\tilde{\theta}_{1n}^c$ can be written compactly as

$$P_{\tilde{\theta}_{1n}^c} = \Phi\left(\frac{-\theta_1^0}{\sigma\sqrt{\omega_{11}}}\right)\epsilon_{(0)} + \frac{1}{\sigma\sqrt{\omega_{11}}}\phi\left(\frac{\tilde{\theta}_1 - \theta_1^0}{\sigma\sqrt{\omega_{11}}}\right)\lambda^+,$$

where $\epsilon_{(0)}$ denotes the unit mass distribution at 0 and λ^+ denotes the Lebesgue measure on $\mathbb{R}^+$.

A similar argument applies to the probability distribution of $\tilde{\lambda}_n$.

(ii) We now consider the probability distribution of $\tilde{\theta}_{2n}^c$. From Remark 21.2 we know that

$$\tilde{\theta}_{2n}^c = (\hat{\theta}_{2n} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\hat{\theta}_{1n}) - (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\tilde{\theta}_{1n}^c.$$

Since $\tilde{\theta}_{1n}^c$ is a function of $\hat{\theta}_{1n}$ and since $\hat{\theta}_{2n} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\hat{\theta}_{1n}$ is independent of $\hat{\theta}_{1n}$, it follows that the two terms in the right-hand side of the preceding equation are mutually independent. Thus the probability distribution of $\tilde{\theta}_{2n}^c$ is the convolution of the probability distribution of $\hat{\theta}_{2n} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\hat{\theta}_{1n}$, which is the normal distribution with mean β_{20} and variance $\sigma^2\omega_{22}^*$, and the probability distribution of $(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'X_1\tilde{\theta}_{1n}^c$, which is a mixture of a discrete distribution and a continuous distribution. It is clearly possible to give an explicit expression for this distribution. The derivation of such an expression is left to the reader. See Exercise 21.2.

e) Asymptotic Distributions

One of the main characteristics of inequality constrained estimators is that their asymptotic distributions are not necessarily normal. To see this, we distinguish two cases according to whether the true parameter value θ^0 satisfies $\theta_1^0 > 0$ or $\theta_1^0 = 0$. To simplify the presentation, we restrict ourselves to the study of $\tilde{\theta}_{1n}^c$.

(i) CASE WHERE $\theta_1^0 > 0$: We know that the unconstrained estimator $\hat{\theta}_{1n}$ converges to the true value $\theta_1^0 > 0$. In particular, $\hat{\theta}_{1n}$ is asymptotically strictly positive with probability one. This implies

$$P(\hat{\theta}_{1n} \leq 0) \to 0,$$

i.e.

$$P(\tilde{\theta}_{1n}^c = \hat{\theta}_{1n}) \to 1.$$

Because the constrained estimator coincides asymptotically with the unconstrained estimator, the constrained estimator will also converge to θ_1^0 and will have the same asymptotic distribution as $\hat{\theta}_{1n}$, i.e., a normal distribution. Specifically, we have

$$\frac{\tilde{\theta}_{1n}^c - \theta_1^0}{\sigma\sqrt{\omega_{11}}} \xrightarrow{d} N(0,1), \tag{21.8}$$

if $\theta_1^0 > 0$. In particular, the asymptotic precision of $\tilde{\theta}_{1n}^c$ is identical to that of $\hat{\theta}_{1n}$.

(ii) CASE WHERE $\theta_1^0 = 0$: The finite sample distribution of $\tilde{\theta}_{1n}^c$ is a mixture of a discrete distribution and a continuous distribution with weights that are independent of the number n of observations

$$P_{\tilde{\theta}_{1n}^c} = \frac{1}{2}\epsilon_{(0)} + \frac{1}{\sigma\sqrt{\omega_{11}}}\phi\left(\frac{\tilde{\theta}_1}{\sigma\sqrt{\omega_{11}}}\right)\lambda^+.$$

It follows that

$$\frac{\tilde{\theta}_{1n}^c}{\sigma\sqrt{\omega_{11}}} \sim \frac{1}{2}\epsilon_{(0)} + \phi(y)\lambda^+, \tag{21.9}$$

if $\theta_1^0 = 0$. Therefore the probability distribution of $\tilde{\theta}_{1n}^c/\sigma\sqrt{\omega_{11}}$ is independent of n. It is equal to the standard normal distribution truncated at zero. Since this probability distribution remains fixed, it is also the asymptotic distribution of the normalized inequality constrained estimator $\tilde{\theta}_{1n}^c/\sigma\sqrt{\omega_{11}}$.

21.2.2 General Results

The same approach as in the preceding subsection can be followed in the general case. Here we shall study simultaneously the main estimation methods, which are maximum likelihood, M-estimation, asymptotic least

square and methods of moments. To do so, we shall use the general framework and notation of Section 10.3.

(i) We consider a criterion function $L_n(\theta)$, which is assumed to satisfy

$$\begin{cases} \frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_0)}{\partial \theta} & \xrightarrow{d} \quad N(0, \mathcal{I}_0), \\ -\frac{1}{n} \frac{\partial^2 L_n(\theta_0)}{\partial \theta \partial \theta'} & \rightarrow \quad \mathcal{J}_0 \text{ with probability 1,} \end{cases} \tag{21.10}$$

for the true parameter value θ_0. To simplify, we shall assume the equality of the matrices $\mathcal{I}_0$ and $\mathcal{J}_0$.

(ii) The true parameter value θ_0 satisfies r inequality constraints, which are written in implicit form as

$$g(\theta) \geq 0 \Leftrightarrow \begin{cases} g_1(\theta) \geq 0, \\ \vdots \\ g_r(\theta) \geq 0, \end{cases} \tag{21.11}$$

where g is a function from $\mathbb{R}^p$ to $\mathbb{R}^r$ whose Jacobian is of rank r.

(iii) The unconstrained estimator $\hat{\theta}_n$ of θ exists and is a solution to the optimization problem $\max_{\theta \in \Theta} L_n(\theta)$. Moreover, when θ_0 belongs to the interior of Θ, the unconstrained estimator $\hat{\theta}_n$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \# \mathcal{J}_0^{-1} \frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_0)}{\partial \theta}.$$

(iv) Lastly, various equality constrained estimators can be considered. Namely, if α is a subset of $\{1, \ldots, r\}$, we can consider the solution to the following constrained optimization problem

$$\begin{cases} \max_\theta L_n(\theta) \\ \text{subject to } g_j(\theta) = 0, \forall j \in \alpha. \end{cases}$$

Such an equality constrained estimator is denoted $\hat{\theta}_n^0(\alpha)$.

a) The Inequality Constrained Estimator

Definition 21.1: *The inequality constrained estimator, which is denoted $\hat{\theta}_n^c$, is a solution to the optimization problem*

$$\begin{cases} \max_\theta L_n(\theta) \\ \text{subject to } g_1(\theta) \geq 0, \ldots, g_r(\theta) \geq 0. \end{cases}$$

The inequality constrained estimator admits different forms depending on which of the inequality constraints are binding.

Example 21.8: To illustrate how the inequality constrained estimator is determined, consider the linear model

$$Y = X_1\theta_1 + X_2\theta_2 + u, \quad u \sim N(0, \sigma^2\mathbf{I}),$$

where θ_1 and θ_2 are scalar parameters constrained to be nonnegative and where X_1 and X_2 are mutually orthogonal vectors.

The method consists in determining first the unconstrained OLS estimator denoted $(\hat{\theta}_{1n}, \hat{\theta}_{2n})'$. If $(\hat{\theta}_{1n}, \hat{\theta}_{2n})'$ is such that $\hat{\theta}_{1n} \geq 0$ and $\hat{\theta}_{2n} \geq 0$, then the unconstrained estimator coincides with $\tilde{\theta}_n^c$.

Otherwise, if $\hat{\theta}_{1n} \leq 0$ or $\hat{\theta}_{2n} \leq 0$, then at least one of the inequality constraints is binding. Suppose that the first inequality constraint is binding. Then we consider the estimator constrained by $\theta_1 = 0$. In the present case the estimator constrained by $\theta_1 = 0$ is $(0, \hat{\theta}_{2n})'$. Thus, if $\hat{\theta}_{2n} \geq 0$, then this estimator coincides with $\tilde{\theta}_n^c$. Otherwise, i.e., if $\hat{\theta}_{2n} < 0$, then the second inequality constraint is also binding.

The preceding argument shows that the inequality constrained estimator is

$$\begin{aligned}\tilde{\theta}_n^c &= \begin{pmatrix} \hat{\theta}_{1n} \\ \hat{\theta}_{2n} \end{pmatrix} \mathbb{I}_{\hat{\theta}_{1n}\geq 0, \hat{\theta}_{2n}\geq 0} + \begin{pmatrix} 0 \\ \hat{\theta}_{2n} \end{pmatrix} \mathbb{I}_{\hat{\theta}_{1n}\leq 0, \hat{\theta}_{2n}\geq 0} \\ &\quad + \begin{pmatrix} \hat{\theta}_{1n} \\ 0 \end{pmatrix} \mathbb{I}_{\hat{\theta}_{1n}\geq 0, \hat{\theta}_{2n}\leq 0} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} \mathbb{I}_{\hat{\theta}_{1n}\leq 0, \hat{\theta}_{2n}\leq 0},\end{aligned}$$

i.e.

$$\tilde{\theta}_n^c = \begin{pmatrix} \hat{\theta}_{1n}\mathbb{I}_{\hat{\theta}_{1n}\geq 0} \\ \hat{\theta}_{2n}\mathbb{I}_{\hat{\theta}_{2n}\geq 0} \end{pmatrix}.$$

b) Kühn–Tücker Multipliers

Alternatively, we can solve the constrained optimization problem by introducing the Kühn–Tücker multipliers. This time, there are r such multipliers associated with the r constraints $g_j(\theta) \geq 0$, $j = 1, \ldots, r$. They are denoted λ_j.

The Lagrangian function is

$$\mathcal{L}_n = L_n(\theta) + \sum_{j=1}^{r} \lambda_j g_j(\theta).$$

The first-order conditions are

$$\begin{cases} \dfrac{\partial \mathcal{L}_n(\tilde{\theta}_n^c, \tilde{\lambda}_n)}{\partial \theta} = 0 \Leftrightarrow \dfrac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta} + \dfrac{\partial g'(\tilde{\theta}_n^c)}{\partial \theta} \tilde{\lambda}_n = 0 \\ \text{sign constraints: } g_j(\tilde{\theta}_n^c) \geq 0, \tilde{\lambda}_{jn} \geq 0, \forall j = 1, \ldots, r, \\ \text{exclusion constraints: } \tilde{\lambda}_{jn} g_j(\tilde{\theta}_n^c) = 0, \forall j = 1, \ldots, r. \end{cases} \tag{21.12}$$

Compared to the Lagrangian function in the OLS case, it is important to note the sign change in the present Lagrangian function. This is because the objective function must be maximized here, while it was minimized in the case of OLS estimation.

Since the Jacobian of g is of rank r, we obtain immediately the relationship between the Kühn–Tücker multipliers and the score vector.

Property 21.1: *we have*

$$\tilde{\lambda}_n = -\left(\frac{\partial g(\tilde{\theta}_n^c)}{\partial \theta'} \Omega_0 \frac{\partial g'(\tilde{\theta}_n^c)}{\partial \theta} \right)^{-1} \frac{\partial g(\tilde{\theta}_n^c)}{\partial \theta'} \Omega_0 \frac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta'},$$

where Ω_0 is an arbitrary positive definite matrix.

c) Asymptotic Properties

We discuss briefly consistency and asymptotic distributions.

(i) Consistency of an estimator obtained as a solution to a maximization of an objective function L_n does not depend on whether there are constraints. Thus, the estimator $\tilde{\theta}_n^c$ is consistent for the true parameter value θ_0.

It follows that

$$\begin{aligned} \frac{1}{n}\tilde{\lambda}_n \quad &\# \quad -\left(\frac{\partial g(\theta_0)}{\partial \theta'} \Omega_0 \frac{\partial g'(\theta_0)}{\partial \theta} \right)^{-1} \frac{\partial g(\theta_0)}{\partial \theta'} \Omega_0 \frac{1}{n} \frac{\partial L_n(\theta_0)}{\partial \theta} \\ &\# \quad -\left(\frac{\partial g(\theta_0)}{\partial \theta'} \Omega_0 \frac{\partial g'(\theta_n)}{\partial \theta} \right)^{-1} \frac{\partial g(\theta_0)}{\partial \theta'} \Omega_0 \, p\lim \frac{1}{n} \frac{\partial L_n(\theta_0)}{\partial \theta}. \end{aligned}$$

The usual asymptotic identification condition on θ_0 implies that

$$\text{plim} \frac{1}{n} \frac{\partial L_n(\theta_0)}{\partial \theta} = 0.$$

Hence $(1/n)\tilde{\lambda}_n \# 0$.

(ii) The study of the asymptotic distributions of $\sqrt{n}(\tilde{\theta}_n^c - \theta_0)$ and $(\sqrt{n}/n)\tilde{\lambda}_n = (1/\sqrt{n})\tilde{\lambda}_n$ is much more difficult. If the true parameter value θ_0 belongs to the interior of the parameter space, i.e., if $g_j(\theta_0) > 0$, $j = 1, \ldots, r$, then $\tilde{\theta}_n^c$ coincides asymptotically with the unconstrained estimator $\hat{\theta}_n$, in which case it is asymptotically normally distributed. On the other hand, if θ_0 belongs to the frontier of the parameter space, the asymptotic distribution of $\tilde{\theta}_n^c$ is much more complex. In the next section we shall determine the asymptotic distributions of some functions of $\tilde{\theta}_n^c$, when θ_0 satisfies $g_j(\theta) = 0$, $j = 1, \ldots, r$.

21.3 Tests of $H_0 : \{g(\theta) = 0\}$ versus $H_1 : \{g(\theta) \geq 0$ and $g(\theta) \neq 0\}$

In this section we shall develop various testing procedures when the parameters are constrained by some inequalities under the maintained or general hypothesis. Our procedures are defined by analogy with the equality constraint case studied in Chapters 17 and 18. This leads to the introduction of Wald tests, Kühn–Tücker multipliers tests, score tests, tests based on the objective function, etc.

We begin with the definitions of the various tests and we shall establish their asymptotic equivalences. Then we shall propose other test statistics that are asymptotically equivalent to the preceding ones and that are easier to compute. Lastly, in Section 21.3.3 we shall determine the asymptotic distribution of these test statistics under the null hypothesis H_0. Their common asymptotic distribution is a mixture of chi-square distributions with degrees of freedom less or equal to the number of constraints.

21.3.1 The Main Test Statistics

We suppose that the objective function $L_n(\theta)$ satisfies condition (21.10) and that the associated matrices $\mathcal{I}_0$ and $\mathcal{J}_0$ are identical.

Let $\tilde{\theta}_n^c$ and $\hat{\theta}_n^0$ denote the inequality constrained estimator and the equality constrained estimator respectively.

Then $\tilde{\theta}_n^c$ is the solution to the optimization problem

$$\begin{cases} \max_\theta L_n(\theta) \\ \text{subject to } g(\theta) \geq 0. \end{cases} \tag{21.13}$$

Let $\tilde{\lambda}_n$ denote the vector of Kühn–Tücker multipliers associated with the inequality constraints $g(\theta) \geq 0$. Let $\partial L_n(\tilde{\theta}_n^c)/\partial\theta$ denote the score vector evaluated at $\tilde{\theta}_n^c$.

On the other hand, let $\hat{\theta}_n^0$ be a solution to the optimization problem

$$\begin{cases} \max_\theta L_n(\theta) \\ \\ \text{subject to } g(\theta) = 0. \end{cases} \tag{21.14}$$

Let $\hat{\lambda}_n^0$ denote the vector of Lagrange multipliers associated with the equality constraints $g(\theta) = 0$, and $\partial L_n(\hat{\theta}_n^0)/\partial\theta$ the score vector evaluated at $\hat{\theta}_n^0$.

a) Test Statistics Based on the Optimal Values of the Objective Function

By analogy with Chapters 17 and 18, we introduce the following definition.

Definition 21.2: *The likelihood ratio type test statistic is*

$$\begin{aligned} \xi_n^R &= 2(L_n(\tilde{\theta}_n^c) - L_n(\hat{\theta}_n^0)) \\ &= 2\left(\max_{\theta:g(\theta)\geq 0} L_n(\theta) - \max_{\theta:g(\theta)=0} L_n(\theta)\right). \end{aligned}$$

Under the null hypothesis H_0: $g(\theta) = 0$, the two estimators $\tilde{\theta}_n^c$ and $\hat{\theta}_n^0$ are consistent for the true value θ_0. Then, using a Taylor series expansion of the statistic ξ_n^R, we obtain the following asymptotic equivalence

$$\begin{aligned} \xi_n^R \quad \# \quad & 2\left(\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta'}\sqrt{n}(\tilde{\theta}_n^c - \theta_0) - \frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial\theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0)\right. \\ & \left. -\frac{1}{2}n(\tilde{\theta}_n^c - \theta_0)'\mathcal{I}_0(\tilde{\theta}_n^c - \theta_0) + \frac{1}{2}n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}_0(\hat{\theta}_n^0 - \theta_0)\right). \end{aligned}$$

We have implicitly assumed that L_n is well defined in a neighborhood of θ_0. This is assumed hereafter.

The first-order conditions of the Lagrangian function give

$$\frac{\partial L_n(\tilde{\theta}_n^c)}{\partial\theta} + \frac{\partial g'(\tilde{\theta}_n^c)}{\partial\theta}\tilde{\lambda}_n = 0.$$

Using a Taylor series expansion around θ_0, we obtain

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}_0\sqrt{n}(\tilde{\theta}_n^c - \theta_0) + \frac{\partial g'(\theta_0)}{\partial \theta}\frac{\tilde{\lambda}_n}{\sqrt{n}} \# 0. \tag{21.15}$$

Similarly, for $\hat{\theta}_n^0$ we obtain

$$\frac{1}{\sqrt{n}}\frac{\partial L_n(\theta_0)}{\partial \theta} - \mathcal{I}_0\sqrt{n}(\hat{\theta}_n^0 - \theta_0) + \frac{\partial g'(\theta_0)}{\partial \theta}\frac{\hat{\lambda}_n^0}{\sqrt{n}} \# 0. \tag{21.16}$$

Substituting these expressions in the Taylor series expansion of the test statistic ξ_n^R obtained earlier, we obtain

$$\begin{aligned}\xi_n^R \quad \# \quad & n(\tilde{\theta}_n^c - \theta_0)'\mathcal{I}_0(\tilde{\theta}_n^c - \theta_0) - n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}_0(\hat{\theta}_n^0 - \theta_0) \\ & +2\left\{-\frac{\tilde{\lambda}_n'}{\sqrt{n}}\frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\tilde{\theta}_n^c - \theta_0) + \frac{\hat{\lambda}_n^{0\prime}}{\sqrt{n}}\frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0)\right\}.\end{aligned}$$

Now, the exclusion condition is

$$g_j(\tilde{\theta}_n^c)\tilde{\lambda}_{jn} = 0, \quad \forall j.$$

Noting that $g_j(\theta_0) = 0$, its first-order Taylor series expansion under H_0 gives

$$\frac{\partial g_j(\theta_0)}{\partial \theta'}\sqrt{n}(\tilde{\theta}_n^c - \theta_0)\frac{\tilde{\lambda}_{jn}}{\sqrt{n}} \# 0, \quad \forall j.$$

Thus

$$\frac{\tilde{\lambda}_n'}{\sqrt{n}}\frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\tilde{\theta}_n^c - \theta_0) \# 0.$$

On the other hand, we have $g(\hat{\theta}_n^0) = g(\theta_0) = 0$. Therefore

$$\frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \# 0.$$

It follows that the test statistic ξ_n^R satisfies

$$\xi_n^R \ \# \ n(\tilde{\theta}_n^c - \theta_0)'\mathcal{I}_0(\tilde{\theta}_n^c - \theta_0) - n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}_0(\hat{\theta}_n^0 - \theta_0). \tag{21.17}$$

b) Hausman–Wald Type Test Statistic

The basic idea underlying such a test statistic is to compare the two estimators $\tilde{\theta}_n^c$ and $\hat{\theta}_n^0$. If we retain the same quadratic form as the one used

to construct the usual Wald test statistic, then we obtain the following definition.

Definition 21.3: *The Hausman–Wald type test statistic is*

$$\xi_n^H = n(\tilde{\theta}_n^c - \hat{\theta}_n^0)\hat{\mathcal{I}}(\tilde{\theta}_n^c - \hat{\theta}_n^0),$$

where $\hat{\mathcal{I}}$ is a consistent estimator of $\mathcal{I}_0$.

The test statistic ξ_n^H is directly related to the test statistic ξ_n^R introduced previously as the next property shows.

Property 21.2: *The test statistics ξ_n^H and ξ_n^R are asymptotically equivalent under the null hypothesis H_0.*

PROOF: Replacing $\tilde{\theta}_n^c - \theta_0$ by $\tilde{\theta}_n^c - \hat{\theta}_n^0 + \hat{\theta}_n^0 - \theta_0$ in the asymptotically equivalent expression (21.17) derived for ξ_n^R, we obtain

$$\xi_n^R \ \# \ n(\tilde{\theta}_n^c - \hat{\theta}_n^0)'\mathcal{I}_0(\tilde{\theta}_n^c - \hat{\theta}_n^0) + 2n(\tilde{\theta}_n^c - \hat{\theta}_n^0)'\mathcal{I}_0(\hat{\theta}_n^0 - \theta_0).$$

On the other hand, it follows from (21.15) and (21.16) that

$$\mathcal{I}_0\sqrt{n}(\tilde{\theta}_n^c - \hat{\theta}_n^0) \ \# \ \frac{\partial g'(\theta_0)}{\partial \theta}\frac{\tilde{\lambda}_n - \hat{\lambda}_n^0}{\sqrt{n}}. \tag{21.18}$$

Hence

$$\xi_n^R \ \# \ n(\tilde{\theta}_n^c - \hat{\theta}_n^0)'\mathcal{I}_0(\tilde{\theta}_n^c - \hat{\theta}_n^0) + \frac{2}{\sqrt{n}}(\tilde{\lambda}_n - \hat{\lambda}_n^0)'\frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0).$$

But the second term is negligible since

$$\frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \ \# \ 0.$$

Therefore

$$\xi_n^R \ \# \ n(\tilde{\theta}_n^c - \hat{\theta}_n^0)'\mathcal{I}_0(\tilde{\theta}_n^c - \hat{\theta}_n^0).$$

□

c) Kühn–Tücker Multiplier Test Statistic

In the proof of Property 21.2 we have established the following approximate relation between the estimator difference and the multiplier difference

$$\mathcal{I}_0\sqrt{n}(\tilde{\theta}_n^c - \hat{\theta}_n^0) \ \# \ \frac{\partial g'(\theta_0)}{\partial \theta}\frac{\tilde{\lambda}_n - \hat{\lambda}_n^0}{\sqrt{n}}.$$

Hence

$$\begin{aligned} \xi_n^H &= n(\tilde{\theta}_n^c - \hat{\theta}_n^0)'\hat{\mathcal{I}}(\tilde{\theta}_n^c - \hat{\theta}_n^0) \\ &\# \quad \frac{1}{n}(\tilde{\lambda}_n - \hat{\lambda}_n^0)'\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}(\tilde{\lambda}_n - \hat{\lambda}_n^0). \end{aligned}$$

It follows that the Hausman type test statistic and the likelihood ratio type statistic (by Property 21.2) can be interpreted asymptotically as a quadratic form based on the difference between multipliers. Specifically, we have the following property.

Property 21.3: *Let be the Kühn–Tücker multiplier test statistic*

$$\xi_n^{KT} = \frac{1}{n}(\tilde{\lambda}_n - \hat{\lambda}_n^0)'\frac{\partial \hat{g}}{\partial \theta'}\hat{\mathcal{I}}^{-1}\frac{\partial \hat{g}'}{\partial \theta}(\tilde{\lambda}_n - \hat{\lambda}_n^0),$$

where $\hat{\mathcal{I}}$ and $\partial \hat{g}/\partial \theta'$ are consistent estimators of $\mathcal{I}_0$ and $\partial g'(\theta_0)/\partial \theta$ respectively, under the null hypothesis H_0. Then ξ_n^{KT} is asymptotically equivalent to ξ_n^R and ξ_n^H under the null hypothesis H_0.

d) Wald Test Statistic

An asymptotically equivalent test statistic can be defined using directly the values of the constraints $g(.)$ under the maintained hypothesis.

Property 21.4: *The Wald test statistic defined by*

$$\xi_n^W = ng'(\tilde{\theta}_n^c)\left(\frac{\partial g(\tilde{\theta}_n^c)}{\partial \theta'}\hat{\mathcal{I}}^{-1}\frac{\partial g'(\tilde{\theta}_n^c)}{\partial \theta}\right)^{-1} g(\tilde{\theta}_n^c)$$

is asymptotically equivalent to ξ_n^R, ξ_n^H, and ξ_n^{KT} under the null hypothesis H_0.

PROOF: We have

$$\begin{aligned} \sqrt{n}g(\tilde{\theta}_n^c) &= \sqrt{n}\left(g(\tilde{\theta}_n^c) - g(\hat{\theta}_n^0)\right) \\ &\# \quad \sqrt{n}g(\theta_0) + \frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\tilde{\theta}_n^c - \theta_0) - \sqrt{n}\left(g(\theta_0) + \frac{\partial g(\theta_0)}{\partial \theta'}(\hat{\theta}_n^0 - \theta_0)\right) \\ &= \frac{\partial g(\theta_0)}{\partial \theta'}\sqrt{n}(\tilde{\theta}_n^c - \hat{\theta}_n^0) \\ &\# \quad \left(\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\right)\frac{\tilde{\lambda}_n - \hat{\lambda}_n^0}{\sqrt{n}}, \end{aligned}$$

where the last equality follows from equation (21.18). The desired result follows immediately from the expressions for the test statistics ξ_n^{KT} and ξ_n^W. □

Remark 21.3: In particular, if $\mathcal{I}_0 = \mathbf{I}$ (the identity matrix) and if the null hypothesis is of the form H_0: $\theta = 0$, then the Wald test statistic reduces to

$$\xi_n^W = n\|\tilde{\theta}_n^c\|^2.$$

e) Score test

Lastly, we can use Property 21.1 with $\boldsymbol{\Omega}_0 = \mathcal{I}_0^{-1}$ so as to obtain

$$\frac{1}{\sqrt{n}}\tilde{\lambda}_n = -\left(\frac{\partial g(\tilde{\theta}_n^c)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{\partial g'(\tilde{\theta}_n^c)}{\partial \theta}\right)^{-1}\frac{\partial g(\tilde{\theta}_n^c)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{1}{\sqrt{n}}\frac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta},$$

i.e.

$$\frac{1}{\sqrt{n}}\tilde{\lambda}_n \# -\left(\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{1}{\sqrt{n}}\frac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta}.$$

A similar relation between $(1/\sqrt{n})\hat{\lambda}_n^0$ and $(1/\sqrt{n})\,\partial L_n(\hat{\theta}_n^0)/\partial\theta$ can also be derived. Therefore we obtain

$$\begin{aligned}\xi_n^{KT} \# & \frac{1}{\sqrt{n}}(\tilde{\lambda}_n - \hat{\lambda}_n^0)'\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\frac{1}{\sqrt{n}}(\tilde{\lambda}_n - \hat{\lambda}_n^0) \\ \# & \frac{1}{\sqrt{n}}\left(\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} - \frac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta}\right)'\mathcal{I}_0^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\left(\frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1} \\ & \times \frac{\partial g(\theta_0)}{\partial \theta'}\mathcal{I}_0^{-1}\frac{1}{\sqrt{n}}\left(\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} - \frac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta}\right).\end{aligned}$$

Thus we have the next property.

Property 21.5: *Let*

$$\begin{aligned}\xi_n^S = & \frac{1}{n}\left(\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} - \frac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta}\right)'\hat{\mathcal{I}}^{-1}\frac{\partial \hat{g}'}{\partial \theta} \\ & \times \left(\frac{\partial \hat{g}}{\partial \theta'}\hat{\mathcal{I}}^{-1}\frac{\partial \hat{g}'}{\partial \theta}\right)^{-1}\frac{\partial \hat{g}}{\partial \theta'}\hat{\mathcal{I}}^{-1}\left(\frac{\partial L_n(\hat{\theta}_n^0)}{\partial \theta} - \frac{\partial L_n(\tilde{\theta}_n^c)}{\partial \theta}\right)\end{aligned}$$

be the score statistic. Then ξ_n^S *is asymptotically equivalent to the test statistics previously defined, i.e.*

$$\xi_n^S \ \# \ \xi_n^{KT} \ \# \ \xi_n^H \ \# \ \xi_n^W \ \# \ \xi_n^R.$$

21.3.2 Asymptotically Equivalent Optimization Problems

In practice the preceding test statistics may be difficult to compute because they all require to solve an optimization problem subject to some inequality constraints. This optimization problem is all the more difficult to solve as the functions L_n and g are quite complex. It may be therefore useful to replace the latter functions by other functions that are easier to work with but that lead to asymptotically equivalent test statistics.

a) Replacing the Objective Function

As in Property 18.7, we can replace the objective function $L_n(\theta)$ by the function

$$L_n^*(\theta) = -\frac{n}{2}(\hat{\theta}_n - \theta)'\hat{\mathcal{I}}(\hat{\theta}_n - \theta),$$

where $\hat{\theta}_n$ is the unconstrained estimator of θ. Then we can define the following two optimization problems

$$\mathcal{P}^* \begin{cases} \max_\theta L_n^*(\theta) \\ \\ \text{subject to } g(\theta) \geq 0, \end{cases} \qquad \mathcal{P}_0^* \begin{cases} \max_\theta L_n^*(\theta) \\ \\ \text{subject to } g(\theta) = 0. \end{cases}$$

This leads to the following property.

Property 21.6: *The test statistics* ξ_n^{*R}, ξ_n^{*H}, ξ_n^{*KT}, ξ_n^{*W}, *and* ξ_n^{*S} *that are based on the optimization problems* $\mathcal{P}_0^*$ *and* $\mathcal{P}^*$ *are asymptotically equivalent to the test statistics* ξ_n^R, ξ_n^H, ξ_n^{KT}, ξ_n^W, *and* ξ_n^S *under the null hypothesis* H_0.

b) Replacing the Inequality Constraints and Using Duality

We can linearize the specified inequality constraints around the true parameter value. This leads to the following quadratic optimization

problems subject to linear constraints

$$\bar{\mathcal{P}}: \begin{cases} \max_\theta -\frac{n}{2}(\hat{\theta}_n - \theta)'\hat{\mathcal{I}}(\hat{\theta}_n - \theta) \\ \\ \text{subject to } g(\theta_0) + \dfrac{\partial g(\theta_0)}{\partial \theta'}(\theta - \theta_0) \geq 0, \end{cases}$$

and

$$\bar{\mathcal{P}}_0: \begin{cases} \max_\theta -\frac{n}{2}(\hat{\theta}_n - \theta)'\hat{\mathcal{I}}(\hat{\theta}_n - \theta) \\ \\ \text{subject to } g(\theta_0) + \dfrac{\partial g(\theta_0)}{\partial \theta'}(\theta - \theta_0) = 0. \end{cases}$$

These two optimization problems cannot be used directly since they both involve the true parameter value θ_0. However, we note that a random variable that is asymptotically equivalent to our preceding test statistics is given by the optimal value of the objective function in the optimization problem

$$\mathcal{P}: \begin{cases} \max_\theta -n(\hat{\theta}_n - \theta)'\hat{\mathcal{I}}(\hat{\theta}_n - \theta) + n(\hat{\theta}_n - \bar{\theta}_n^0)'\hat{\mathcal{I}}(\hat{\theta}_n - \bar{\theta}_n^0) \\ \\ \text{subject to } g(\theta_0) + \dfrac{\partial g(\theta_0)}{\partial \theta'}(\theta - \theta_0) \geq 0, \end{cases}$$

where $\bar{\theta}_n^0$ is the solution to the optimization problem $\bar{\mathcal{P}}_0$. This is because the optimal value of the objective function in the optimization problem $\mathcal{P}$ is

$$2(\max_{\bar{\mathcal{P}}} L_n^*(\theta) - \max_{\bar{\mathcal{P}}_0} L_n^*(\theta)).$$

Then, using the theory of duality for quadratic optimizations under linear constraints, we see that the preceding optimal value is also the optimal value of the objective function in the optimization problem

$$\mathcal{D}: \begin{cases} \min_\lambda \frac{1}{n}(\lambda - \bar{\lambda}_n^0)' \dfrac{\partial g(\theta_0)}{\partial \theta'} \hat{\mathcal{I}}^{-1} \dfrac{\partial g'(\theta_0)}{\partial \theta}(\lambda - \bar{\lambda}_n^0) \\ \\ \text{subject to } \lambda \geq 0, \end{cases}$$

where $\bar{\lambda}_n^0$ is the vector of Lagrange multipliers in the optimization problem $\bar{\mathcal{P}}_0$.

Lastly, we can replace the matrix associated with the quadratic form defining the objective function by a consistent estimator under the null hypothesis H_0 without modifying the asymptotic results. Similarly, we can replace $\bar{\lambda}_n^0$ by $\hat{\lambda}_n^0$, which is the vector of Lagrange multipliers associated with the equality constrained optimization problem (21.14).

Hence we have proved the following property.

Property 21.7: *Consider the optimization problem*

$$\begin{cases} \min_\lambda \frac{1}{n}(\lambda - \hat{\lambda}_n^0)' \frac{\partial g(\hat{\theta}_n^0)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n^0)^{-1} \frac{\partial g'(\hat{\theta}_n^0)}{\partial \theta} (\lambda - \hat{\lambda}_n^0) \\ \\ \text{subject to } \lambda \geq 0. \end{cases}$$

Let ξ_n^D denote the optimal value of the objective function. This test statistic, which is called the dual test statistic, is asymptotically equivalent to ξ_n^R, ξ_n^H, ξ_n^{KT}, ξ_n^W, and ξ_n^S under the null hypothesis H_0.

The dual test statistic ξ_n^D only requires (i) the computation of the equality constrained estimator $(\hat{\theta}_n^0, \hat{\lambda}_n^0)$ and (ii) a quadratic maximization under linear (nonnegativity) constraints on the vector λ. Moreover, we note that the dimension of λ, which is equal to the number of constraints, is typically much lower than the dimension of θ.

21.3.3 Asymptotic Distribution of the Test Statistics Under the Null Hypothesis H_0

It remains to derive the asymptotic distributions of the various test statistics defined in Sections 21.3.1 and 21.3.2 under the null hypothesis H_0: $g(\theta) = 0$. Since these test statistics are all asymptotically equivalent, it suffices to find the asymptotic distribution of only one of them.

We shall use the dual test statistic of Property 21.7. Then, using Property 18.2 with $\mathcal{I}_0 = \mathcal{J}_0$ and $\mathbf{M} = 0$, we note that

$$Var\left(\frac{1}{\sqrt{n}}\hat{\lambda}_n^0\right) = \Omega_0^{-1} = \left(\frac{\partial g(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}.$$

Hence the problem reduces to finding the asymptotic distribution of

$$\xi = \min_{\lambda : \lambda \geq 0} \frac{1}{n}(\lambda - \hat{\lambda}_n^0)\Omega_0(\lambda - \hat{\lambda}_n^0).$$

Let

$$\hat{\mu}_n^0 = \Omega_0^{1/2} \frac{\hat{\lambda}_n^0}{\sqrt{n}}, \quad \mu = \Omega_0^{1/2} \frac{\lambda}{\sqrt{n}}.$$

Then it is equivalent to find the asymptotic distribution of

$$\xi = \min_{\{\mu : \Omega_0^{-1/2}\mu \geq 0\}} \|\mu - \hat{\mu}_n^0\|^2,$$

where $\hat{\mu}_n^0$ is asymptotically distributed as a standard normal variable.

Lemma 21.1: *Let X be a standard normal random vector of dimension r, i.e., $X \sim N(0, \mathbf{I})$. Let $\mathbf{R}$ be a nonsingular symmetric matrix of size r. Then the distribution of*

$$\xi = \min_{x:\mathbf{R}x \geq 0} \|X - x\|^2$$

is a mixture of chi-square distributions. Namely

$$\xi \sim \sum_{d=0}^{r} \omega(d)\chi^2(d),$$

where $\omega(d)$, $d = 0, \ldots, r$, is a sequence of weights satisfying

$$\omega(d) \geq 0 \quad \textit{and} \quad \sum_{d=0}^{r} \omega(d) = 1,$$

and where $\chi^2(0)$ denotes the point mass distribution at zero.

To simplify the statement of Lemma 21.1, we provide some explicit expressions for the weights $\omega(d)$ only at the end of its proof given next. For the same reason, the proof is given for the case $r = 2$ only. We give, however, a sketch of the proof in the general case afterwards.

PROOF: The proof consists of five steps.

STEP 1: The constraints $\mathbf{R}x \geq 0$ bear upon the components of the vector $\mathbf{R}x$. Since $\mathbf{R}$ is a symmetric matric, these constraints can be written as $R_1'x \geq 0$ and $R_2'x \geq 0$, where R_1 and R_2 denote the columns of the matrix $\mathbf{R}$. Then the set of points $\{x : R_1'x \geq 0, R_2'x \geq 0\}$ constitutes a convex cone $C_{(12)}$ in $\mathbb{R}^2$.

To determine such a cone, it suffices to find two vectors $\tilde{R}_1$ and $\tilde{R}_2$ such that

$$\tilde{R}_1'R_1 > 0, \quad \tilde{R}_2'R_2 > 0, \quad \tilde{R}_1'R_2 = 0, \quad \tilde{R}_2'R_1 = 0.$$

See Figure 21.1. Thus the cone $C_{(12)}$ is

$$C_{(12)} = \{x = \lambda_1\tilde{R}_1 + \lambda_2\tilde{R}_2, \text{with } \lambda_1 \geq 0, \lambda_2 \geq 0\}.$$

Note that there exist many choices for the vectors $\tilde{R}_1$ and $\tilde{R}_2$, which are defined up to some strictly positive multiplicative constants. Among

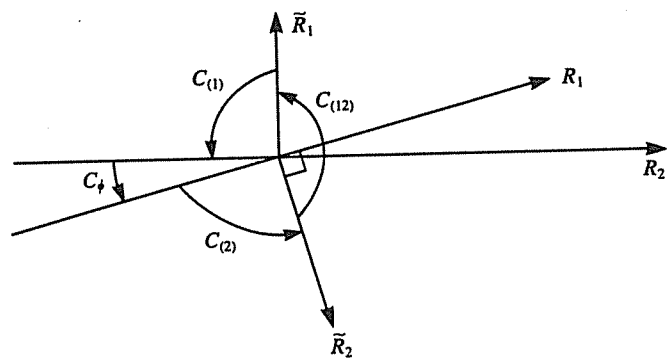


Figure 21.1: Cone $C_{(12)}$

these choices, we can consider the column vectors of the inverse of the matrix $\mathbf{R}$, i.e.

$$\tilde{\mathbf{R}} = (\tilde{R}_1, \tilde{R}_2) = \mathbf{R}^{-1}.$$

This is because we have

$$\tilde{R}_1' R_1 = 1 > 0, \quad \tilde{R}_2' R_2 = 1 > 0, \quad \tilde{R}_1' R_2 = 0, \quad \tilde{R}_2' R_1 = 0.$$

STEP 2: We introduce the following convex cones

$$C_{(1)} = \{x : \tilde{R}_1' x \geq 0, \tilde{R}_2' x \leq 0\} = \{x = \lambda_1 \tilde{R}_1 + \lambda_2 R_2, \lambda_1 \geq 0, \lambda_2 \leq 0\},$$
$$C_{(2)} = \{x : R_1' x \leq 0, \tilde{R}_2' x \geq 0\} = \{x = \lambda_1 R_1 + \lambda_2 \tilde{R}_2, \lambda_1 \leq 0, \lambda_2 \geq 0\},$$
$$C_\emptyset = \{x : \tilde{R}_1' x \leq 0, \tilde{R}_2' x \leq 0\} = \{x = \lambda_1 R_1 + \lambda_2 R_2, \lambda_1 \leq 0, \lambda_2 \leq 0\}.$$

STEP 3: The optimal value ξ of the objective function is nothing else other than

$$\xi = \|X - Proj_{C_{(12)}} X\|^2 = \|Proj_{C_\emptyset}\|^2,$$

where $Proj_C X$ denotes the projection of X on the cone C. Thus Lemma 21.1 can be viewed as a generalization of the Cochran Theorem to projections on some convex sets, which are not necessarily subspaces. The projection $Proj_{C_\emptyset} X$ can take various forms according to the location of X. Specifically

$$\begin{aligned}
&\text{if } X \in C_{(12)} &&: \|Proj_{C_\emptyset} X\|^2 = 0,\\
&\text{if } X \in C_{(1)} &&: \|Proj_{C_\emptyset} X\|^2 = \|Proj_{R_2} X\|^2,\\
&\text{if } X \in C_{(2)} &&: \|Proj_{C_\emptyset} X\|^2 = \|Proj_{R_1} X\|^2,\\
&\text{if } X \in C_\emptyset &&: \|Proj_{C_\emptyset} X\|^2 = \|X\|^2.
\end{aligned}$$

Thus the projection of X on the cone $C_\emptyset$ coincides with the projection on a subspace of dimension 0 if $X \in C_{(12)}$, the projection on a subspace of dimension 1 if $X \in C_{(1)} \cup C_{(2)}$, and the projection on a subspace of dimension 2 if $X \in C_\emptyset$. See Figure 21.2.

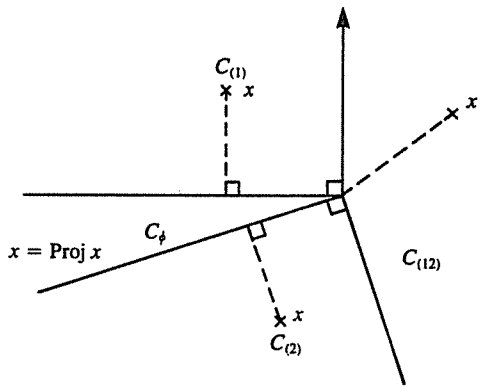


Figure 21.2: Projections on the Cone $C_\emptyset$

STEP 4: It remains to determine the probability distribution of ξ, i.e., to determine $P(\xi \in A)$, where A is an arbitrary subset. We have

$$\begin{aligned} P(\xi \in A) &= P(\xi \in A \mid X \in C_{(12)})P(X \in C_{(12)}) \\ &\quad +P(\xi \in A \mid X \in C_{(1)})P(X \in C_{(1)}) \\ &\quad +P(\xi \in A \mid X \in C_{(2)})P(X \in C_{(2)}) \\ &\quad +P(\xi \in A \mid X \in C_\emptyset)P(X \in C_\emptyset). \end{aligned}$$

We also note that the random variables $X_1^2+X_2^2$ and X_2/X_1 are mutually independent since X is distributed as $N(0, \mathbf{I})$,

Now, the event $X \in C_\emptyset$ depends only on the variable X_2/X_1. It follows that

$$\begin{aligned} P(\xi \in A \mid X \in C_\emptyset) &= P(X_1^2 + X_2^2 \in A \mid X \in C_\emptyset) \\ &= P(X_1^2 + X_2^2 \in A) \\ &= P_2(A), \end{aligned}$$

where P_2 denotes the probability distribution of a chi-square random variable with two degrees of freedom.

Next, we have

$$P(\xi \in A \mid X \in C_{(12)}) = P(0 \in A \mid X \in C_{(12)}) = P_0(A),$$

where P_0 denotes the point mass distribution at zero.

It remains to compute the terms of the form $P(\xi \in A \mid X \in C_{(1)})$ and $P(\xi \in A \mid X \in C_{(2)})$. In fact, since the standard normal distribution is invariant under orthogonal transformations, we can always assume that each of the cones $C_{(1)}$ and $C_{(2)}$, which have an orthogonal angle at their vertex, coincides with the positive orthant. Then we must compute

$$\begin{aligned} P(X_1^2 \in A \mid X_1 \geq 0, X_2 \geq 0) &= P(X_1^2 \in A \mid X_1 \geq 0) \\ &= P(X_1^2 \in A) \\ &= P_1(A), \end{aligned}$$

where P_1 denotes the probability distribution of a chi-square random variable with one degree of freedom. The first equality follows by independence of X_1 and X_2 and the second equality follows by symmetry of the distribution of X_1.

STEP 5: Lastly, letting

$$\begin{aligned} \omega(0) &= P(X \in C_{(12)}), \\ \omega(1) &= P(X \in C_{(1)}) + P(X \in C_{(2)}), \\ \omega(2) &= P(X \in C_{\emptyset}), \end{aligned}$$

we obtain

$$\xi \sim \omega(0)\chi^2(0) + \omega(1)\chi^2(1) + \omega(2)\chi^2(2)$$

as desired. □

SKETCH OF THE PROOF IN THE GENERAL CASE: When r is arbitrary, the proof follows the same five steps. In particular, it is necessary to introduce the column vectors R_j, $j = 1, \ldots, r$, of the matrix $\mathbf{R}$ and the column vectors $\tilde{R}_j$, $j = 1, \ldots, r$, of its inverse $\mathbf{R}^{-1}$. Then, for every subset A of $\{1, \ldots, r\}$, we define the cone

$$\begin{aligned} C_A = \Bigg\{ & x = \sum_{j \in A} \lambda_j \tilde{R}_j + \sum_{j \notin A} \lambda_j R_j, \\ & \text{with } \lambda_j \geq 0 \text{ if } j \in A, \text{ and } \lambda_j \leq 0 \text{ if } j \notin A \Bigg\}. \end{aligned}$$

The desired result is

$$\xi \sim \sum_{d=0}^{r} \omega(d)\chi^2(d),$$

where

$$\omega(d) = \sum_{\{A:\ \#A=r-d\}} P(X \in C_A). \tag{21.19}$$

□

Applying Lemma 21.1 to our problem, we obtain immediately the following property.

Property 21.8: *Under the null hypothesis H_0, the test statistics ξ_n^R, ξ_n^H, ξ_n^W, ξ_n^{KT}, ξ_n^S, and ξ_n^D are all asymptotically distributed as a mixture of chi-square distributions, namely as*

$$\sum_{d=0}^{r} \omega(d)\chi^2(d),$$

where the weights are given by equation (21.19) and the matrix $\mathbf{R}$ *is*

$$\mathbf{R} = \left(\frac{\partial g(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial g'(\theta_0)}{\partial \theta} \right)^{-1/2}.$$

In practice the weights must be estimated since they depend on θ_0, which is unknown.

Remark 21.4: When $r = 2$, the weights can be explicitly determined. From the proof, we know that these weights are obtained from probabilities of the form $P(X \in C_A)$. Because the distribution of X is invariant under orthogonal transformations, this probability is equal to $\alpha_A/2\pi$, where α_A denotes the angle at the vertex of the cone C_A. It follows at once that $\omega(1) = 1/2$.

Moreover, we have

$$\cos \alpha_{12} = \frac{\tilde{R}_1' \tilde{R}_2}{\sqrt{\tilde{R}_1' \tilde{R}_1}\ \sqrt{\tilde{R}_2' \tilde{R}_2}}.$$

Since $\mathbf{R}^{-1}$ is a symmetric matrix, this ratio is equal to

$$\frac{r_{12}}{\sqrt{r_{11}}\sqrt{r_{22}}},$$

where r_{ij} is the generic term of the matrix

$$\mathbf{R}^{-2} = \frac{\partial g(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial g'(\theta_0)}{\partial \theta}.$$

Thus we obtain

$$\cos \alpha_{12} = \frac{\dfrac{\partial g_1(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \dfrac{\partial g_2'(\theta_0)}{\partial \theta}}{\sqrt{\dfrac{\partial g_1(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \dfrac{\partial g_1'(\theta_0)}{\partial \theta}} \sqrt{\dfrac{\partial g_2(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \dfrac{\partial g_2'(\theta_0)}{\partial \theta}}}.$$

Hence

$$\omega(0) = \frac{\alpha_{12}}{2\pi}, \quad \omega(2) = \frac{1}{2} - \frac{\alpha_{12}}{2\pi}.$$

Therefore the mixture of chi-square distributions is

$$\omega(0)\chi^2(0) + \frac{1}{2}\chi^2(1) + \left(\frac{1}{2} - \omega(0)\right)\chi^2(2),$$

where $\omega(0)$ can be computed from

$$\omega(0) = \frac{1}{2\pi} \arccos(\rho_{as}(g_1(\hat{\theta}_n), g_2(\hat{\theta}_n))) \tag{21.20}$$

and ρ_{as} denotes the asymptotic correlation coefficient. Thus the weight $\omega(0)$ is directly related to the asymptotic correlation coefficient between the estimated constraints $g_1(\hat{\theta}_n)$ and $g_2(\hat{\theta}_n)$.

21.3.4 Examples

a) A Bidimensional Example

Suppose that we are interested in a problem where the parameter θ is of dimension 2. We also assume that the problem has been normalized so that the unconstrained maximum likelihood estimator $\hat{\theta}_n$ of θ satisfies $\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, \sigma^2\mathbf{I})$. For instance, this is the case in the linear model $y = \mathbf{X}\theta + u$, $u \sim N(0, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is a $n \times 2$ matrix satisfying $\mathbf{X}'\mathbf{X} = \mathbf{I}$.

We impose the inequalities

$$\begin{cases} \theta^1 \geq 0, \\ \theta^2 - r\theta^1 \geq 0. \end{cases}$$

We want to test the null hypothesis H_0: $\theta = 0$.

An estimator, which is asymptotically equivalent to the inequality constrained maximum likelihood estimator $\tilde{\theta}_n^c$, can be obtained by considering the orthogonal projection $\tilde{\theta}_n$ of $\hat{\theta}_n$ on to the set defined by the above inequalities. See Figure 21.3.

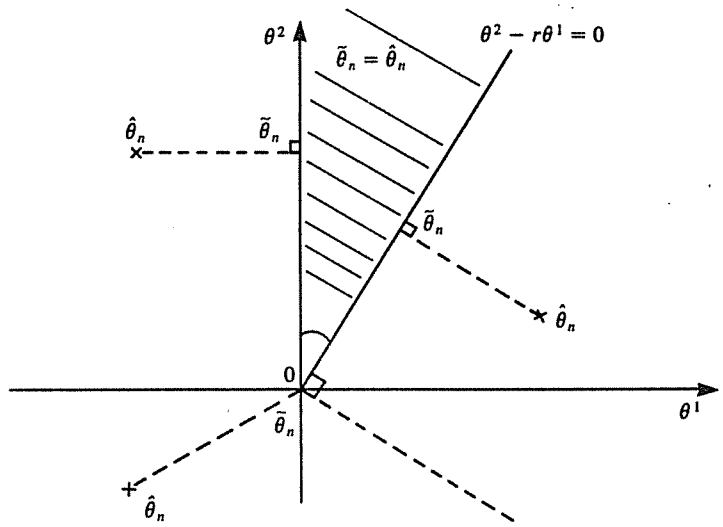


Figure 21.3: Projections of $\hat{\theta}_n$

Thus the critical region of the Hausman–Wald test is

$$W = \left\{ n\frac{\|\tilde{\theta}_n\|^2}{\hat{\sigma}_n^2} \leq c \right\},$$

where c is the $(1-\alpha)$-th quantile of the mixture of chi-square distributions

$$\left(\frac{1}{2} - \frac{\phi}{2\pi}\right)\chi^2(0) + \frac{1}{2}\chi^2(1) + \frac{\phi}{2\pi}\chi^2(2).$$

Here ϕ denotes the angle of the straight lines given by $\theta^1 = 0$ and $\theta^2 - r\theta^1 = 0$, while $\hat{\sigma}_n^2$ is a consistent estimator of σ^2. Therefore c is determined by

$$\frac{1}{2}P(\chi^2(1) > c) + \frac{\phi}{2\pi}P(\chi^2(2) > c) = \alpha.$$

The shape of the critical region W is illustrated in Figure 21.4.

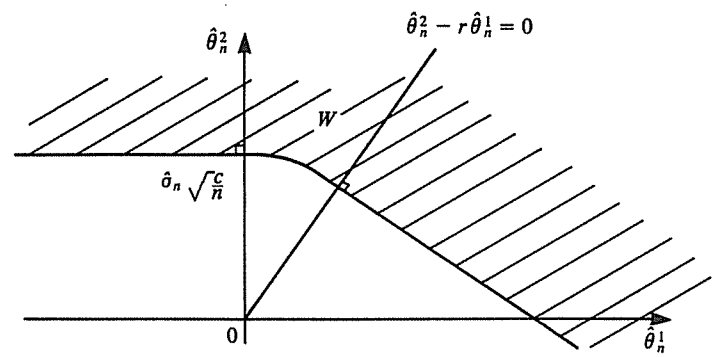


Figure 21.4: Critical Region of the Hausman–Wald Test

b) A Panel Data Model

We consider the linear model

$$y_{it} = x_{it}\beta + u_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T,$$

where the observations are indexed by individual i and by time t. We assume that the error u_{it} can be decomposed as

$$u_{it} = \mu_i + \nu_{it},$$

where the variables μ_i and ν_{it} are mutually independent and distributed as $N(0, \sigma_\mu^2)$ and $N(0, \sigma_\nu^2)$ respectively. The component μ_i is interpreted as an individual effect.

The variance covariance matrix of the errors is not diagonal. Ordering the observations first with respect to i and then with respect to t, this matrix is

$$\mathbf{V}u = \sigma_\nu^2 \mathbf{I}_{NT} + \sigma_\mu^2 \mathbf{I}_N \otimes e_T e_T',$$

where e_T denotes the vector of dimension T whose components are all equal to one and where $\mathbf{I}_{NT}$ (say) denotes the identity matrix of size $NT \times NT$.

Since

$$(\mathbf{V}u)^{-1} = \frac{1}{\sigma_\nu^2}\mathbf{I}_{NT} - \frac{\sigma_\mu^2}{\sigma_\nu^2(\sigma_\nu^2 + T\sigma_\mu^2)}\mathbf{I}_N \otimes e_T e_T' \equiv \mathbf{\Sigma}^{-1},$$

the log-likelihood function is

$$L_n = -\frac{NT}{2}\log 2\pi - \frac{1}{2}\log\det\mathbf{\Sigma} - \frac{1}{2}(y - \mathbf{X}\beta)'\mathbf{\Sigma}^{-1}(y - \mathbf{X}\beta),$$

where n is the total number of observations, i.e., $n = NT$. Hereafter, we assume that N increases to infinity while T remains fixed.

We wish to test for the absence of individual effects. Thus the null hypothesis is H_0: $\sigma_\mu^2 = 0$, while the maintained hypothesis is H: $\sigma_\mu^2 \geq 0$.

(i) The Lagrange multiplier associated with the constraint $\sigma_\mu^2 = 0$ is equal to

$$\begin{aligned}\hat{\lambda}_n^0 &= \frac{\partial L_n(\hat{\theta}_n^0)}{\partial \sigma_\mu^2} \\ &= \left(\frac{1}{2}\left(-\frac{NT}{\sigma_\nu^2 + T\sigma_\mu^2} + \frac{1}{(\sigma_\nu^2 + T\sigma_\mu^2)^2} u'(\mathbf{I}_N \otimes e_T e_T')u\right)\right)_{\theta=\hat{\theta}_n^0} \\ &= -\frac{1}{2}\frac{NT}{\hat{\sigma}_\nu^2} + \frac{1}{2}\frac{1}{\hat{\sigma}_\nu^4}(y - \mathbf{X}\hat{\beta})'(\mathbf{I}_N \otimes e_T e_T')(y - \mathbf{X}\hat{\beta}),\end{aligned}$$

where $\hat{\beta}$ denotes the OLS estimator of β and $\hat{\sigma}_\nu^2$ denotes the constrained estimator of σ_ν^2, i.e.

$$\hat{\sigma}_\nu^2 = \frac{(y - \mathbf{X}\hat{\beta})'(y - \mathbf{X}\hat{\beta})}{NT} = \frac{\hat{u}'\hat{u}}{NT}.$$

Then, expressing the Lagrange multiplier in terms of the residuals $\hat{u}$, we obtain

$$\hat{\lambda}_n^0 = \frac{(NT)^2}{2(\hat{u}'\hat{u})^2}(\hat{u}'(\mathbf{I}_N \otimes e_T e_T')\hat{u} - \hat{u}'\hat{u}).$$

(ii) To compute the dual test statistic ξ_n^D, it is also necessary to evaluate the quantity

$$\frac{1}{NT}\frac{\partial g(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\hat{\theta}_n^0)^{-1}\frac{\partial g'(\hat{\theta}_n^0)}{\partial \theta}.$$

In the present case such a quantity is asymptotically equivalent to the element of the matrix $(-\partial^2 L_n/\partial\theta\partial\theta')^{-1}$ corresponding to σ_μ^2 evaluated at $\hat{\theta}_n^0$. Because the matrix $\partial^2 L_n/\partial\theta\partial\theta'$ is block diagonal with blocks corresponding to β and $(\sigma_\nu^2, \sigma_\mu^2)$, it is easy to show that the desired element is

$$\frac{2(\hat{u}'\hat{u})^2}{(NT)^3(T-1)}.$$

(iii) Then the optimization problem of Property 21.7 becomes

$$\xi_n^D = \min_{\lambda \geq 0} \frac{2(\hat{u}'\hat{u})^2}{(NT)^3(T-1)}(\lambda - \hat{\lambda}_n^0)^2.$$

21.3. TESTS OF $H_0 : \{g(\theta) = 0\}$ VS. $H_1 : \{g(\theta) \geq 0$ AND $g(\theta) \neq 0\}$

Two cases must be distinguished:
Case 1: if $\hat{\lambda}_n^0 \geq 0$, i.e., if $\hat{u}'(\mathbf{I}_N \otimes e_T e_T')\hat{u} \geq \hat{u}'\hat{u}$, then

$$\xi_n^D = 0.$$

Case 2: if $\hat{\lambda}_n^0 < 0$, i.e., if $\hat{u}'(\mathbf{I}_N \otimes e_T e_T')\hat{u} < \hat{u}'\hat{u}$, then

$$\begin{aligned}
\xi_n^D &= \frac{2(\hat{u}'\hat{u})^2}{(NT)^3(T-1)}(\hat{\lambda}_n^0)^2 \\
&= \frac{2(\hat{u}'\hat{u})^2}{(NT)^3(T-1)} \frac{(NT)^4}{4(\hat{u}'\hat{u})^4}(\hat{u}'\hat{u})^2 \left(1 - \frac{\hat{u}'(\mathbf{I}_N \otimes e_T e_T')\hat{u}}{\hat{u}'\hat{u}}\right)^2 \\
&= \frac{NT}{2(T-1)} \left(1 - \frac{\hat{u}'(\mathbf{I}_N \otimes e_T e_T')\hat{u}}{\hat{u}'\hat{u}}\right)^2.
\end{aligned}$$

(iv) Under H_0, ξ_n^D is asymptotically distributed as $(1/2)\chi^2(0) + (1/2)\chi^2(1)$, where $\chi^2(0)$ is the mass point distribution at 0 (see Exercise 21.4). It follows that the critical region of the corresponding test is

$$W = \{\xi_n^D \geq \chi^2_{1-2\alpha}(1)\}$$

for an asymptotic level equal to α with $\alpha < 0.5$.

(v) It is interesting to note that the expression for the dual test statistic emphasizes the role played by the difference

$$A_n = \hat{u}'(\mathbf{I}_N \otimes e_T e_T')\hat{u} - \hat{u}'\hat{u}.$$

Developing this difference gives

$$A_n = \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{u}_{it}\right)^2 - \sum_{i=1}^{N}\sum_{t=1}^{T} \hat{u}_{it}^2.$$

Then, as $N \to \infty$ we can see that

$$\frac{1}{(T-1)NT} A_n = \frac{1}{(T-1)NT} \left(\sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{u}_{it}\right)^2 - \sum_{i=1}^{N}\sum_{t=1}^{T} \hat{u}_{it}^2\right)$$

converges to

$$\operatorname{plim} \frac{1}{(T-1)NT} A_n = \frac{1}{(T-1)T} \left(E\left(\sum_{t=1}^{T} u_{it}\right)^2 - E\left(\sum_{t=1}^{T} u_{it}^2\right)\right)$$

$$= \frac{1}{(T-1)T}\left(E\left(T\mu_i + \sum_{t=1}^{T}\nu_{it}\right)^2 - TE(u_{it}^2)\right)$$
$$= \frac{1}{(T-1)T}(T^2\sigma_\mu^2 + T\sigma_\nu^2 - T(\sigma_\mu^2 + \sigma_\nu^2)) = \sigma_\mu^2.$$

Thus, the statistic $(1/(T-1)NT)A_n$ is a consistent estimator of σ_μ^2. This provides a Wald type interpretation to the dual test statistic ξ_n^D.

21.4 Tests of $H_0 : \{g(\theta) \geq 0\}$

21.4.1 Some General Remarks

Testing situations, where the null hypothesis is of the form H_0: $g(\theta) \geq 0$, arise when, for instance, one wishes to test that some matrices are positive semidefinite. Then the null hypothesis H_0 corresponds to some inequality constraints on the parameters, while the maintained hypothesis H corresponds to the situation where the parameters are left unconstrained.

One can construct many measures of discrepancy between the two hypotheses H_0 and H. For instance, let $\hat{\theta}_n$ and $\tilde{\theta}_n^c$ denote respectively the unconstrained estimator and the inequality constrained estimator of the parameter θ. Then one can introduce a statistic based on the difference between the optimal values of the unconstrained and constrained objective functions, i.e.

$$\xi_n^R = 2(L_n(\hat{\theta}_n) - L_n(\tilde{\theta}_n^c)).$$

Alternatively, one can propose a Hausman–Wald type statistic of the form

$$\xi_n^W = \min_{\{\theta: g(\theta)\geq 0\}} n(\hat{\theta}_n - \theta)'\mathcal{I}(\hat{\theta}_n)(\hat{\theta}_n - \theta).$$

It is, however, important to note that the present testing problem is quite different from the one studied in Section 21.3. For the null hypothesis is here composite with a nonempty interior. This leads to two remarks:

1. It is now more difficult to propose some tests based on a local approach and, in particular, to propose some Lagrange multiplier type tests or score type tests. This is because the neighborhoods of some points under the null hypothesis are entirely included in the null hypothesis.

2. Under the null hypothesis H_0, the asymptotic distribution of ξ_n^W, if it exists, will have different forms according to the true parameter value θ_0, which satisfies $g(\theta_0) \geq 0$. If one considers critical regions of the form $W = \{\xi_n^W \geq c\}$, then the condition on the risk of Type I error α is

$$\sup_{\theta_0 : g(\theta_0) \geq 0} P_{\theta_0}(\xi_n^W \geq c) \leq \alpha.$$

21.4.2 A Special Case

The construction of a test can be achieved under additional assumptions. Below, we present heuristically the basic idea underlying such a construction, which consists of three steps.

STEP 1: First, we note that the present testing problem can be viewed as a testing problem of the parameter $\beta \equiv g(\theta)$. In general, information on β provided by $\hat{\theta}_n$ is contained in $\hat{\beta}_n \equiv g(\hat{\theta}_n)$. Moreover, the estimator $\hat{\theta}_n$ is asymptotically normally distributed. Thus, if g is differentiable, then $\hat{\beta}_n$ is asymptotically normally distributed and we have

$$\sqrt{n}(\hat{\beta}_n - g(\theta_0)) \sim N\left(0, \frac{\partial g(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial g'(\theta_0)}{\partial \theta}\right).$$

Then an asymptotically equivalent method for testing the null hypothesis H_0 consists in determining the optimal value, denoted $\tilde{\xi}_n^W$, of the objective function in the optimization problem

$$\begin{cases} \min_\beta n(\hat{\beta}_n - \beta)' \hat{\mathbf{V}}_n^{-1} (\hat{\beta}_n - \beta) \\ \text{subject to } \beta \geq 0, \end{cases}$$

where $\hat{\mathbf{V}}_n$ is a consistent estimator of the asymptotic variance covariance matrix of $\hat{\beta}_n$, i.e., of $V_{as}(\sqrt{n}(\hat{\beta}_n - g(\theta_0))) = \mathbf{V}(\theta_0) \equiv \mathbf{V}_0$.

STEP 2: It remains to determine the asymptotic distribution(s) of $\tilde{\xi}_n^W$ under the null hypothesis H_0.

First, we note that the asymptotic distribution of $\tilde{\xi}_n^W$ is equal to that of $\bar{\xi}_n^W$, where $\bar{\xi}_n^W$ is the optimal value of the objective function in the optimization problem

$$\begin{cases} \min_\beta n(\hat{\gamma}_n - \gamma)'(\hat{\gamma}_n - \gamma) \\ \text{subject to } \mathbf{V}(\theta_0)^{1/2}\gamma \geq 0, \end{cases}$$

where $\hat{\gamma}_n = \hat{\mathbf{V}}_n^{-1/2}\hat{\beta}_n$. Then we have $V_{as}(\sqrt{n}(\hat{\gamma}_n - \gamma_0)) = \mathbf{I}$, where $\gamma_0 = \mathbf{V}_0^{-1/2}g(\theta_0)$.

The statistic $\bar{\xi}_n^W$ can be interpreted as n times the squared length of the vector $\hat{\gamma}_n - Proj_C\hat{\gamma}_n$, where C denotes the convex cone $C = \{\gamma : \mathbf{V}(\theta_0)^{1/2}\gamma \geq 0\}$. This is formally stated in the following property.

Property 21.9: *We have*

$$\begin{aligned}\bar{\xi}_n^W &= n\|\hat{\gamma}_n - Proj_C\hat{\gamma}_n\|^2 \\ &= n(\|\hat{\gamma}_n\|^2 - \|Proj_C\hat{\gamma}_n\|^2) \\ &= n\|Proj_{C^0}\|^2,\end{aligned}$$

where C^0 denotes the polar cone of C.

This property follows from the definition of a polar cone and from Pythagora's Theorem applied to projections on the cones. It is easily interpreted geometrically as Figure 21.5 illustrates.

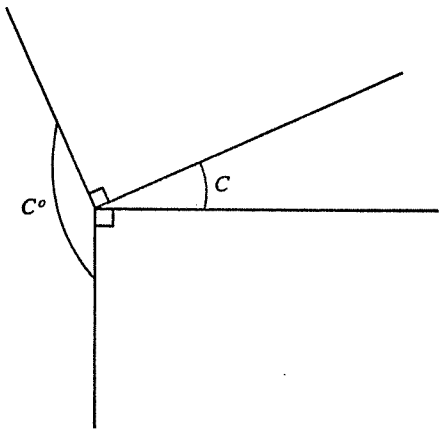


Figure 21.5: Cone and Polar Cone

The equality $\bar{\xi}_n^W = n\|Proj_{C^0}\hat{\gamma}_n\|^2$ leads to the following corollary.

Corollary 21.1: *The "statistic" $\bar{\xi}_n^W$ satisfies*

$$\bar{\xi}_n^W = n\tilde{\tilde{\gamma}}_n^{C'}\tilde{\tilde{\gamma}}_n^{C},$$

where $\tilde{\tilde{\gamma}}_n^c$ *is the solution to the optimization problem*

$$\begin{cases} \min_\gamma n(\hat{\gamma}_n - \gamma)'(\hat{\gamma}_n - \gamma) \\ \text{subject to } \gamma \in C^0. \end{cases}$$

Thus this statistic coincides with a Wald statistic for testing the null hypothesis

$$\tilde{H}_0 : \{\gamma_0 = 0\} \quad vs. \quad \tilde{H}_1 : \{\gamma_0 \in C^0, \gamma_0 \neq 0\}$$

where $\gamma_0 \equiv g(\theta_0) = 0$ *(see Remark 21.3).*

Therefore the test of H_0: $\gamma_0 \in C$ against the maintained (unconstrained) hypothesis H and the test of $\tilde{H}_0$: $\gamma_0 = 0$ against $\tilde{H}$: $\gamma_0 \in C^0$ lead to the same test statistic. See Figure 21.6.

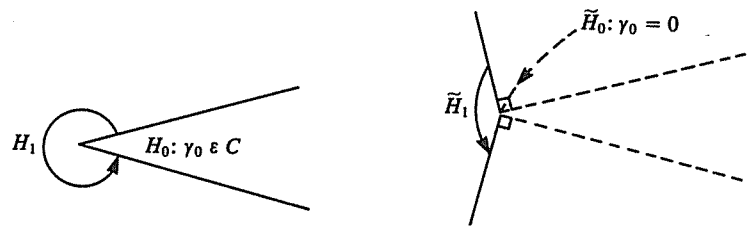


Figure 21.6: Hypotheses Leading to the Same Test Statistic

Intuitively, the part of H_1 that is the most "orthogonal" to H_0 is $\tilde{H}_1$: $\gamma_0 \in C^0$. Thus testing H_0: $\gamma_0 \in C$ reduces to examining whether the model can be usefully modified in this "orthogonal direction."

STEP 3: We now give a condition that enables us to solve the problem mentioned earlier concerning the risk of Type I error.

Property 21.10: *Consider a parameterization of the model such that the functions* β *of the parameter* θ *that are eventually constrained appear directly as* $\theta = (\alpha, \beta)'$. *If*

$$\mathcal{I}(\theta) = \begin{pmatrix} \mathcal{I}_{\alpha\alpha} & 0 \\ 0 & \mathcal{I}_{\beta\beta} \end{pmatrix}$$

and if the matrix $\mathbf{V}(\theta)$ *depends on the parameter* θ *only through* α, *then*

$$\sup_{\theta_0 : g(\theta_0) \geq 0} P_{\theta_0}(\xi_n^W \geq c) = \sup_{\theta_0 : g(\theta_0) = 0} P_{\theta_0}(\xi_n^W \geq c).$$

PROOF: Under the null hypothesis H_0: $g(\theta_0) \geq 0$, i.e., H_0: $\beta_0 \geq 0$, i.e., H_0: $\gamma_0 \in C$, we have asymptotically $\sqrt{n}(\hat{\gamma}_n - \gamma_0) \sim N(0, \mathbf{I})$, where $\gamma_0 \in C$.

Then we apply the following lemma whose proof is easy and left to the reader.

Lemma: *If $\gamma_0 \in C$ and if u is an arbitrary vector, then*

$$\|Proj_{C^0}(u + \gamma_0)\|^2 \leq \|Proj_{C^0} u\|^2.$$

Now, because $\mathbf{V}(\theta_0)$ is independent of γ_0, the cone C is independent of γ_0. Moreover, for every value of the parameters α_0 and β_0 with $\beta_0 \geq 0$, we have $\gamma_0 \in C$ and $\sqrt{n}\hat{\gamma}_n = Z + \sqrt{n}\gamma_0$, where Z is asymptotically distributed as a standard normal random variable. It follows that

$$\begin{aligned} P_{\alpha_0,\beta_0}(\bar{\xi}_n^W \geq c) &= P_{\alpha_0,\beta_0}(n\|Proj_{C^0}\hat{\gamma}_n\|^2 \geq c) \\ &= P_{\alpha_0}(Proj_{C^0}\|(Z + \sqrt{n}\hat{\gamma}_0)\|^2 \geq c). \end{aligned}$$

Then, using the inequality of the above lemma, we obtain

$$P_{\alpha_0,\beta_0}(\bar{\xi}_n^W \geq c) \leq P_{\alpha_0}(Proj_{C^0}\|Z\|^2 \geq c) = P_{\alpha_0,0}(\bar{\xi}_n^W \geq c).$$

This establishes the desired result. □

We can now state the main result of this section.

Property 21.11: *Under the same conditions as in Property 21.10, for any α, $0 < \alpha < 1$, there exists a test with asymptotic level α for testing the null hypothesis H_0: $g(\theta_0) \geq 0$. This test is given by a critical region of the form $W = \{\xi_n^W \geq c\}$. The critical level c is determined by the condition*

$$\mathcal{Q}[-\infty, c] = 1 - \alpha,$$

where $\mathcal{Q}$ denotes the probability distribution of the following mixture of chi-square distributions

$$\mathcal{Q} = \sum_{d=0}^{r} \omega(r-d)\chi^2(d),$$

with $\omega(d)$ being the weights obtained for the test of $\tilde{H}_0$: $\gamma_0 = 0$ against $\tilde{H}_1$: $\{\gamma_0 \in C^0, \gamma_0 \neq 0\}$.

PROOF: This is a direct consequence of Corollary 21.1. The test statistic ξ_n^W can be interpreted as a statistic for testing $\tilde{H}_0$ against $\tilde{H}_1$. From Property 21.10, it suffices to determine its asymptotic distribution under the null hypothesis $\tilde{H}_0$: $\gamma_0 = 0$, i.e., H_0: $g(\theta_0) = 0$. It only suffices to remember that the cone C must be replaced by its polar cone C^0. This replacement modifies accordingly the weights $\omega(d)$. □

21.4.3 An Example

We consider the linear model

$$Y = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + u, \quad u \sim N(0, \sigma^2\mathbf{I}),$$

where β_1 and β_2 are possibly vectors. We wish to test the null hypothesis H_0: $\beta_1 \geq 0$.

Without loss of generality, we can assume that the column vectors of $\mathbf{X}_1$ are orthogonal to the column vectors of $\mathbf{X}_2$. Then the information matrix is block diagonal, namely

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\beta_1\beta_1} & 0 & 0 \\ 0 & \mathcal{I}_{\beta_2\beta_2} & 0 \\ 0 & 0 & \mathcal{I}_{\sigma^2\sigma^2} \end{pmatrix}.$$

Moreover, if $\hat{\beta}_{1n}$ denotes the ordinary least squares estimator of β_1, then we have

$$V(\hat{\beta}_{1n}) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} = \mathbf{V}(\theta),$$

which is a matrix that does not depend on the parameter β_1. Hence the conditions given in Properties 21.10 and 21.11 are satisfied.

It follows that a statistic for testing H_0 is given by the optimal value of the objective function in the optimization problem

$$\begin{cases} \min_{\beta_1} (\hat{\beta}_{1n} - \beta_1)' \dfrac{\mathbf{X}_1'\mathbf{X}_1}{\hat{\sigma}_n^2} (\hat{\beta}_{1n} - \beta_1) \\ \\ \text{subject to } \beta_1 \geq 0, \end{cases}$$

where $\hat{\sigma}_n^2$ is a consistent estimator of the error variance. Then the critical level is determined by using the probability distribution of such a test statistic when $\beta_1 = 0$. When β_1 is a scalar, this probability distribution is the mixture of chi-squares distributions $\frac{1}{2}\chi^2(0) + \frac{1}{2}\chi^2(1)$.

21.5 Exercises

EXERCISE 21.1:

a) Determine the characteristic polynomial associated with the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Verify that the discriminant of this polynomial is $\Delta = (a-d)^2+4bc$.

b) When $\Delta < 0$, write the stability condition in terms of a, b, c, and d.

c) When $\Delta > 0$, determine the roots of the deterministic polynomial. Then derive the stability conditions in terms of a, b, c, and d.

EXERCISE 21.2:

a) Consider two independent random variables X and Y such that $X \sim N(m, \sigma^2)$ and Y is distributed as $\Phi(-\theta/\sigma)\epsilon(0) + (1/\sigma)\phi((y-\theta)/\sigma)\lambda^+$. Determine the probability density of $Z = X + Y$.

b) Apply a) to the estimator $\tilde{\theta}^c_{2n}$ obtained in Section 21.2.1.

EXERCISE 21.3: When the number of constraints is equal to 1, verify that the test statistics ξ^R_n, ξ^H_n, ξ^W_n, ξ^{KT}_n, and ξ^D_n are asymptotically distributed as $\frac{1}{2}\chi^2(0) + \frac{1}{2}\chi^2(1)$.

EXERCISE 21.4: Consider the asymptotic distribution of the test statistics considered in this chapter, namely

$$\xi \sim \sum_{d=0}^{r} \omega(d)\chi^2(d).$$

Show that the sum of the weights $\omega(d)$ corresponding to the odd values of d is always equal to 1/2. Show that the same property holds for the even values of d.

EXERCISE 21.5: Consider the linear model

$$Y = \mathbf{X}\beta + u, \quad u \sim N(0, \mathbf{I}),$$

where the parameter β is constrained by the *scalar* inequality $R\beta \geq r$. Find the Wald statistic for testing the null hypothesis H_0: $R\beta = r$.

Verify that this test coincides with the one-sided UMP test of this null hypothesis.

EXERCISE 21.6: Consider a linear model with linear inequality constraints. Can the Kühn–Tücker statistic for testing H_0: $\mathbf{R}\beta = r$ be interpreted in terms of the coefficient of determination?

EXERCISE 21.7: Consider the following disequilibrium model with endogenous price adjustments defined by

$$\begin{cases} D_t = a_1 P_t + x_{1t} b_1 + u_{1t}, & a_1 < 0, \\ S_t = a_2 P_t + x_{2t} b_2 + u_{2t}, & a_2 > 0, \\ P_t - P_{t-1} = \lambda (D_t - S_t), & \lambda > 0, \\ Q_t = \min(D_t, S_t). \end{cases}$$

The endogenous variables are Q_t and P_t.

a) Why is the equilibrium hypothesis written as $\lambda = +\infty$?

b) Suppose that the errors u_{1t} and u_{2t} are mutually independent and distributed as $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively. Write the likelihood function of the model. Then find the limit of this likelihood function as λ increases to $+\infty$. Verify that the limit likelihood function corresponds to the likelihood function of the equilibrium model.

c) Test the equilibrium hypothesis by noting that such a test is a test of

$$H_0 : \bar{\lambda} = 0 \quad \text{vs.} \quad H_1 : \bar{\lambda} > 0,$$

where $\bar{\lambda} \equiv 1/\lambda$.

21.6 References

Bartholomew, D.J. (1961). "A Test of Homogeneity of Means under Restricted Alternatives," *Journal of the Royal Statistical Society*, B, 23, 239–281.

Gourieroux, C., Holly A., and Monfort A. (1982). "Likelihood Ratio Test, Wald Test and Kühn–Tücker Test in Linear Models with Inequality Constraints on the Regression Parameters," *Econometrica*, 50, 63–80.

Hillier, G.H. (1986). "Joint Tests for Zero Restrictions on Non Negative Regression Coefficients," *Biometrika*, 73, 657–669.

Nuesh, P.E. (1966). "On the Problem of Testing Location in Multivariate Problems for Restricted Alternatives," *The Annals of Mathematical Statistics*, 37, 113–119.

Perlman, M.D. (1969). "On Sided Problems in Multivariate Analysis," *The Annals of Mathematical Statistics*, 40, 549–567.

Rodgers, A.J. (1986). "Modified Lagrange Multiplier Tests for Problems with One-Sided Alternatives," *Journal of Econometrics*, 31, 341–361.

Wolak, F.A. (1987). "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model," *Journal of the American Statistical Association*, 82, 782–793.

CHAPTER 22

Nonnested Tests and Model Selection Criteria

In the preceding chapters the problem of choosing a model was addressed through the so-called theory of classical tests. See especially Chapters 14–19 and Chapter 21. In such a theory (applied to the parametric case), a general hypothesis is maintained and is characterized by a family of conditional probability densities

$$\ell(y \mid x; \theta) = \prod_{i=1}^{n} f(y \mid x_i; \theta).$$

This family specifies the probability distribution of the endogenous variables Y_i, $i = 1, \ldots, n$, given the exogenous variables X_i, $i = 1, \ldots, n$. The parameter vector θ is assumed to belong to some parameter space Θ.

The null hypothesis H_0 is then defined by imposing some equality and/or inequality constraints on some functions of the parameter vector θ. As a result, the null hypothesis is naturally included or *nested* in the maintained hypothesis.

In this chapter we shall study the case where the null and alternative hypotheses are no longer naturally nested in a general hypothesis. Two approaches will be discussed. The first approach consists in extending the classical testing procedures (Wald test, score test, etc.) to the case of nonnested hypotheses. In contrast, within the second approach each competing model is evaluated by means of a numerical criterion whose maximization or minimization over the set of competing models will define the selection procedure.

To study these two approaches, we need some concepts of Kullback proximity and pseudo true values. These concepts, which were already introduced in Chapter 1, are generalized to the case of conditional models in the next section.

22.1 Kullback Proximity and Pseudo True Values

22.1.1 Definition

We consider n pairs (Y_i, X_i), $i = 1, \ldots, n$, of independent and identically distributed random vectors. We are interested in the conditional probability distribution of the vectors Y_i, $i = 1, \ldots, n$, given the vectors X_i, $i = 1, \ldots, n$. We assume that this distribution has a probability density with respect to a product measure $\mu^{\otimes n}$

$$\ell_0(y \mid x) = \prod_{i=1}^{n} f_0(y_i \mid x_i),$$

where $y' = (y'_1, \ldots, y'_n)$ and $x' = (x'_1, \ldots, x'_n)$.

Such a conditional probability density is unknown. To evaluate its proximity to a specified parametric model defined by

$$\mathcal{P} = \{\ell(y \mid x; \theta) = \prod_{i=1}^{n} f(y_i \mid x_i; \theta), \theta \in \Theta\},$$

we can consider the concept of Kullback information, which was defined in Section 1.2.3.

Definition 22.1: *The Kullback proximity between the true conditional density $\ell_0(y \mid x)$ and the family of conditional densities $\ell(y \mid x; \theta)$ is the quantity*

$$K_n(\mathcal{P} \mid \ell_0) = \min_{\theta \in \Theta} K_n(\ell(y \mid x; \theta)/\ell_0(y \mid x)),$$

where $K_n(\ell(y \mid x; \theta)/\ell_0(y \mid x))$ is the Kullback information evaluated conditionally given $X = x$

$$K_n(\ell(y \mid x; \theta)/\ell_0(y \mid x)) = \frac{1}{n} E_0 \left(\log \frac{\ell_0(Y \mid x)}{\ell(Y \mid x; \theta)} \right).$$

Definition 22.1 differs from Definition 1.6 in two ways:

(i) The proximity is evaluated conditionally given x so that $K_n(\mathcal{P} \mid \ell_0)$ becomes a function of x. Such a function is nonnegative and is equal to zero if and only if the true conditional probability density ℓ_0 belongs to the family $\mathcal{P}$.

(ii) The above proximity measure has been normalized by dividing by n. Thus $K_n(\ell(y \mid x;\theta)/\ell_0(y \mid x))$ can be naturally interpreted as an empirical mean, namely

$$\begin{aligned} K_n(\ell(y \mid x;\theta)/\ell_0(y \mid x)) &= \frac{1}{n}\sum_{i=1}^{n} E_0\left(\log \frac{f_0(Y_i \mid x_i)}{f(Y_i \mid x_i;\theta)}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_1(f(y \mid x_i;\theta)/f_0(y \mid x_i)). \end{aligned}$$

Definition 22.2: *A parameter value θ_n^* that minimizes the proximity of $K_n(\ell(y \mid x;\theta)/\ell_0(y \mid x))$, or equivalently, that maximizes $E_0 \log \ell(Y \mid x;\theta)$, is called a pseudo true value in finite samples (associated with the value x of the exogenous variables). The conditional probability distribution with density $\ell(y \mid x;\theta_n^*)$ is called the pseudo true density in finite samples.*

The pseudo true value in finite samples θ_n^* depends on the values of the exogenous variables. Thus, in principle, it should be denoted $\theta_n^*(x)$. Before observing x, it must be considered as a random variable $\theta_n^*(X)$. To simplify notation, however, the argument X is omitted hereafter.

The asymptotic properties of θ_n^*, in particular its consistency, can be studied by replacing the optimization problem in finite samples by the corresponding asymptotic optimization problem. From the interpretation of K_n as an empirical mean and the Strong Law of Large Numbers, it follows that

$$K_n(\ell(y \mid x;\theta)/\ell_0(y \mid x)) = \frac{1}{n}\sum_{i=1}^{n} E_0\left(\log \frac{f_0(Y_i \mid x_i)}{f(Y_i \mid x_i;\theta)}\right)$$

converges to

$$K(f(y \mid x;\theta)/f_0(y \mid x)) = E_x E_0\left(\log \frac{f_0(Y_i \mid X_i)}{f(Y_i \mid X_i;\theta)}\right). \tag{22.1}$$

Under suitable regularity conditions such as those given in Chapter 24, the solution θ_n^* of the optimization problem in finite samples converges to a solution θ^* of the limit optimization problem defined by

$$\min_{\theta \in \Theta} K(f(y \mid x;\theta)/f_0(y \mid x)).$$

Definition 22.3: *A solution to the limit optimization problem*

$$\min_{\theta\in\Theta} E_x E_0 \left(\log \frac{f_0(Y \mid X)}{f(Y \mid X;\theta)} \right),$$

or equivalently, to $\max_{\theta\in\Theta} E_x E_0 \log f(Y \mid X;\theta)$, *is called an asymptotic pseudo true value of the parameter* θ.

The optimal value of the objective function, which is

$$K(\mathcal{P} \mid \ell_0) \equiv \min_{\theta\in\Theta} K(f(y|x;\theta)/f_0(y|x)) = K(f(y \mid x;\theta^*)/f_0(y \mid x)),$$

is called the asymptotic Kullback proximity between the true probability distribution ℓ_0 *and the model* $\mathcal{P}$.

The asymptotic pseudo true value is a nonrandom vector. Thus, in general, it differs from the pseudo true value in finite samples. However, in the special case of sampling models, i.e., models without exogenous variables, these two concepts are identical, i.e., $\theta_n^* = \theta^*$.

Remark 22.1: In Chapter 9 we saw that the pseudo maximum likelihood estimator $\hat{\theta}_n$ of the parameter θ is a consistent estimator of the solution to the optimization problem

$$\max_{\theta\in\Theta} \sum_{i=1}^{n} E_x E_0 \log f(Y_i \mid X_i;\theta).$$

That is, $\hat{\theta}_n$ is a consistent estimator of the asymptotic pseudo true value θ^*. Since θ_n^* also converges to this value, it follows that the difference $\hat{\theta}_n - \theta_n^*$ converges to zero.

22.1.2 Pseudo True Values and Proximities

a) Example: The Linear Model

PSEUDO TRUE VALUES IN FINITE SAMPLES:

We assume that the true conditional distribution is the normal distribution $N(m_0, \sigma_0^2\mathbf{I})$ and that the specified model is the Gaussian linear model

$$Y = \mathbf{X}_1 b + u, \quad \text{with } u \sim N(0, \sigma^2\mathbf{I}).$$

From Example 1.16 we know that the pseudo true values in finite samples b_n^* and σ_n^{*2} are given by

$$\begin{cases} \mathbf{X}_1 b_n^* = \mathbf{M}_1 m_0, \\ \sigma_n^{*2} = \sigma_0^2 + \dfrac{\|m_0 - \mathbf{M}_1 m_0\|^2}{n}, \end{cases} \tag{22.2}$$

where $\mathbf{M}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ is the orthogonal projection matrix on to the space generated by the columns of $\mathbf{X}_1$. It follows that

$$b_n^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' m_0.$$

Note that these pseudo true values are functions of the values of the exogenous variables and of the moments of the true conditional distribution.

A SPECIAL CASE:

To study this example further, suppose now that the conditional mean m_0 is of the form $m_0 = \mathbf{X}_2 a_0$, where $\mathbf{X}_2$ is a matrix whose number of columns is independent of the number of observations.

Then we have

$$b_n^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 a_0 = \left(\sum_{i=1}^n x_{i1}' x_{i1}\right)^{-1}\left(\sum_{i=1}^n x_{i1}' x_{i2}\right) a_0$$

and

$$\begin{aligned}\sigma_n^{*2} &= \sigma_0^2 + \frac{\|\mathbf{X}_2 a_0 - \mathbf{M}_1\mathbf{X}_2 a_0\|^2}{n} \\ &= \sigma_0^2 + \frac{1}{n} a_0' \left(\sum_{i=1}^n x_{i2}' x_{i2} - \left(\sum_{i=1}^n x_{i2}' x_{i1}\right)\left(\sum_{i=1}^n x_{i1}' x_{i1}\right)^{-1} \right. \\ &\quad \left. \times \left(\sum_{i=1}^n x_{i1}' x_{i2}\right)\right) a_0,\end{aligned}$$

where x_{i1} and x_{i2} denote the rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively.

ASYMPTOTIC PSEUDO TRUE VALUES:

Suppose that the rows (X_{i1}, X_{i2}), $i = 1, \ldots, n$, are independent and identically distributed. Then it is easy to determine the asymptotic pseudo true values by considering the limits of b_n^* and σ_n^{*2}. We obtain

$$b^* = (E(X_{i1}'X_{i1}))^{-1}E(X_{i1}'X_{i2})a_0 \tag{22.3}$$

$$\begin{aligned}\sigma^{*2} &= \sigma_0^2 + a_0' \left(E(X_{i2}'X_{i2}) - E(X_{i2}'X_{i1})(E(X_{i1}'X_{i1}))^{-1}\right. \\ &\quad \left. \times E(X_{i1}'X_{i2})\right) a_0.\end{aligned} \tag{22.4}$$

The asymptotic pseudo true values depend on the quantities a_0, σ_0^2, and the second-order moments of the true distribution of the exogenous variables. In particular, $\sigma_n^{*2} = \sigma^{*2} = \sigma_0^2$ if X_1 includes the columns of X_2, and $b_n^* = b^* = a_0$ if $X_1 = X_2$.

b) Example: Causality Measures

THE TRUE DISTRIBUTION:

The hypothesis of independence of the observations is clearly not necessary for computing the Kullback proximity measure. This measure can be evaluated in other situations such as when the observations Y_t, $t = 1, \ldots, T$, are generated by a stationary process. This case is considered now.

It is assumed that the true probability distribution that has generated the observations $Y_1, \ldots, Y_T$ corresponds to a bidimensional Gaussian autoregressive process of order 1. Let Y_{1t} and Y_{2t} denote the two components of Y_t. We have

$$\begin{cases} Y_{1,t} &= a_{11}^0 Y_{1,t-1} + a_{12}^0 Y_{2,t-1} + u_{1,t}, \\ Y_{2,t} &= a_{21}^0 Y_{1,t-1} + a_{22}^0 Y_{2,t-1} + u_{2,t}. \end{cases} \tag{22.5}$$

The errors $(u_{1t}, u_{2t}), t = 1, \ldots, T$, are assumed to be independently and identically normally distributed over time with zero means and variance covariance matrix

$$\Sigma_0 = \begin{pmatrix} \sigma_{11}^0 & \sigma_{12}^0 \\ \sigma_{12}^0 & \sigma_{22}^0 \end{pmatrix}.$$

The coefficients a_{ij}^0, $i = 1, 2$, $j = 1, 2$, are such that the process is stable.

Thus the true probability density of $\mathbf{Y} = (Y_0', Y_1', \ldots, Y_T')'$ is of the form

$$\begin{aligned} \ell_0(y) &= f_0(y_0) \prod_{t=1}^{T} f_0(y_{1t}, y_{2t} \mid y_{1,t-1}, y_{2,t-1}) \\ &= f_0(y_0) \prod_{t=1}^{T} f_0(y_{1t} \mid y_{1,t-1}, y_{2,t-1}) f_0(y_{2t} \mid y_{1,t}, y_{1,t-1}, y_{2,t-1}), \end{aligned} \tag{22.6}$$

where $f_0(y_0)$ denotes the true marginal probability density of Y_0.

THE MODEL:

The specified model corresponds to the hypothesis of noncausality in the Granger sense of Y^2 on Y^1. See Definition 1.12. The model is

$$\begin{cases} Y_{1,t} &= a_{11} Y_{1,t-1} + u_{1,t}, \\ Y_{2,t} &= a_{21} Y_{1,t-1} + a_{22} Y_{2,t-1} + u_{2t}, \end{cases} \tag{22.7}$$

where the errors are independently and identically normally distributed over time with zero means and variance covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Thus, the generic probability density for Y in the corresponding family $\mathcal{P}$ is of the form

$$\ell(y;\theta) = f(y_0)\prod_{t=1}^{T} f(y_{1t} \mid y_{1,t-1}) f(y_{2t} \mid y_{1t}, y_{1,t-1}, y_{2,t-1}). \tag{22.8}$$

The marginal probability density of Y_0 is left unconstrained, namely, $Y_0 \sim N(m, \mathbf{S})$ for some vector m and matrix $\mathbf{S}$.

The parameters of the model are $\theta = (m, \mathbf{S}, a_{11}, a_{12}, a_{22}, \sigma_{11}, \sigma_{22}, \sigma_{12})$.

DETERMINATION OF PROXIMITIES:

The Kullback proximity is

$$\begin{aligned} K_T(\mathcal{P} \mid \ell_0) &= \frac{1}{T}\min_{\theta} E_0\left(\log\frac{\ell_0(Y)}{\ell(Y;\theta)}\right) \\ &= \frac{1}{T}\min_{\theta}\left\{E_0 \log\frac{f_0(Y_0)}{f(Y_0)}\right. \\ &\quad + \sum_{t=1}^{T} E_0 \log\frac{f_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1})}{f(Y_{1t} \mid Y_{1,t-1})} \\ &\quad \left. + \sum_{t=1}^{T} E_0 \log\frac{f_0(Y_{2t} \mid Y_{1t}, Y_{1,t-1}, Y_{2,t-1})}{f(Y_{2t} \mid Y_{1,t}, Y_{1,t-1}, Y_{2,t-1})}\right\}. \end{aligned} \tag{22.9}$$

The first term is null by letting $m = E_0 Y_0$ and $\mathbf{S} = V_0 Y_0$. The third term can also be set equal to zero. This is because the parameters of the conditional distribution of Y_{2t} given Y_{1t}, $Y_{1,t-1}$, and $Y_{2,t-1}$ in the family $\mathcal{P}$ are unconstrained and, hence, can be chosen to be equal to the corresponding true values.

It remains to examine the second term. Because the maximum of the quantity $E_0(\log f(Y_{1t} \mid Y_{1,t-1}) \mid Y_{1,t-1})$ is attained when $f(y_{1t} \mid y_{1,t-1}) = f_0(y_{1t} \mid y_{1,t-1})$, it follows that

$$\begin{aligned} &E_0 \log\frac{f_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1})}{f(Y_{1t} \mid Y_{1,t-1})} \\ &\quad = E_0 \log f_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1}) - E_0 E_0(\log f(Y_{1t} \mid Y_{1,t-1}) \mid Y_{1,t-1}) \\ &\quad \geq E_0 \log f_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1}) - E_0 E_0(\log f_0(Y_{1t} \mid Y_{1,t-1}) \mid Y_{1,t-1}). \end{aligned}$$

Collecting results, we obtain

$$K_T(\mathcal{P} \mid \ell_0) = \frac{1}{T}\sum_{t=1}^{T} E_0\left(\log \frac{f_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1})}{f_0(Y_{1t} \mid Y_{1,t-1})}\right).$$

Then, using the stationarity of the data generating process, we obtain

$$K_T(\mathcal{P} \mid \ell_0) = K(\mathcal{P} \mid \ell_0) = E_0\left(\log \frac{f_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1})}{f_0(Y_{1t} \mid Y_{1,t-1})}\right). \quad (22.10)$$

The latter quantity, which is zero if only if Y_2 does not cause Y_1, provides a natural measure of causality.

AN ALTERNATIVE EXPRESSION FOR THE CAUSALITY MEASURE:
Because all the conditional distributions are normal, we have

$$\begin{aligned} K(\mathcal{P} \mid \ell_0) &= E_0 \log\left(\frac{f_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1})}{f_0(Y_{1t} \mid Y_{1,t-1})}\right) \\ &= E_0\left(-\frac{1}{2}\frac{(Y_{1t} - E_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1}))^2}{V_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1})}\right. \\ &\quad +\frac{1}{2}\frac{(Y_{1t} - E_0(Y_{1t} \mid Y_{1,t-1}))^2}{V_0(Y_1 \mid Y_{1,t-1})} \\ &\quad -\frac{1}{2}\log V_0(Y_{1t} \mid Y_{1,t-1}, Y_{2,t-1}) \\ &\quad \left.+\frac{1}{2}\log V_0(Y_{1t} \mid Y_{1,t-1})\right). \end{aligned}$$

Hence

$$K(\mathcal{P} \mid \ell_0) = \frac{1}{2}\log \frac{V_0(Y_{1t} \mid Y_{1,t-1})}{V_0(Y_{1,t} \mid Y_{1,t-1}, Y_{2,t-1})}. \quad (22.11)$$

The latter expression can be generalized in different ways. See Exercise 22.1.

22.2 Tests of Nonnested Hypotheses

22.2.1 Nonnested Hypotheses

We consider a sequence (Y_i, X_i), $i = 1, 2, \ldots$, of independent and identically distributed random vectors. The hypotheses we are interested

in bear on the conditional probability density of Y_i given X_i. These hypotheses correspond to the following two parametric families of conditional densities

$$\begin{cases} H_g & : \{g(y_i \mid x_i; \alpha), \alpha \in A \subset \mathbb{R}^G\}, \\ H_h & : \{h(y_i \mid x_i; \beta), \beta \in B \subset \mathbb{R}^H\}. \end{cases} \tag{22.12}$$

We want to know whether the true conditional probability density $f_0(y_i \mid x_i)$ of Y_i given X_i belongs to H_g or to H_h. To answer such a question, we can let one of the two hypotheses be the null hypothesis. The approach being symmetric in H_g and H_h, we can choose a priori either H_g or H_h to be the null hypothesis.

Tests conducted using successively H_g and H_h as the null hypothesis may provide consistent answers such as the rejection of H_g (or H_h) by both tests. However, answers provided by the two tests may be inconsistent. This arises when both hypotheses H_g and H_h are simultaneously accepted or simultaneously rejected. The simultaneous acceptance of H_g and H_h may indicate a lack of data, while the simultaneous rejection of H_g and H_h may indicate that the general framework of analysis is incorrect. These various situations are summarized in the following table.

When $H_0 = H_g$ $\rightarrow$ When $H_0 = H_h$ $\downarrow$	Accept H_g	Accept H_h
Accept H_g	consistent	inconsistent
Accept H_h	inconsistent	consistent

22.2.2 Generalized Wald Tests

Without loss of generality, we take H_g to be the null hypothesis.

(i) Suppose first that the true probability distribution corresponds to a density f_0 in H_g associated with a value α_0 of the parameter α.

The pseudo true value in finite samples of β is denoted $b_n(\alpha_0)$, while the asymptotic pseudo true value of β is denoted $b(\alpha_0)$. Consider now the maximum likelihood estimators of α and β. These estimators are obtained by maximizing respectively

$$L_n^g(\alpha) = \sum_{i=1}^{n} \log g(y_i \mid x_i; \alpha)$$

and

$$L_n^h(\beta) = \sum_{i=1}^{n} \log h(y_i \mid x_i; \beta).$$

Let $\hat{\alpha}_n$ and $\hat{\beta}_n$ denote these estimators.

Because we have assumed that the true distribution satisfies $f_0(y \mid x) = g(y \mid x; \alpha_0)$, then $\hat{\alpha}_n$ is in fact the usual maximum likelihood estimator of α_0. On the other hand, $\hat{\beta}_n$ is a pseudo maximum likelihood estimator and converges to the asymptotic pseudo true value $b(\alpha_0)$. Therefore the differences $\hat{\beta}_n - b(\hat{\alpha}_n)$ and $\hat{\beta}_n - b_n(\hat{\alpha}_n)$ converge to zero.

(ii) Suppose now that the true probability distribution belongs to H_h and, hence, corresponds to a density with parameter value β_0, i.e.

$$f_0(y \mid x) = h(y \mid x; \beta_0).$$

In terms of convergence of the above estimators we have

$$\lim_n \hat{\alpha}_n = a(\beta_0),$$

which is the pseudo true value of α

$$\lim_n \hat{\beta}_n = \beta_0$$

and

$$\lim_n (\hat{\beta}_n - b(\hat{\alpha}_n)) = \lim_n (\hat{\beta}_n - b_n(\hat{\alpha}_n)) = \beta_0 - b(a(\beta_0)).$$

Note that the latter limit is not equal to zero, in general.

(iii) Thus we can construct a testing procedure based on either one of the two differences $\hat{\beta}_n - b(\hat{\alpha}_n)$ and $\hat{\beta}_n - b_n(\hat{\alpha}_n)$. This can be done using the next property, which gives the asymptotic distributions of $\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))$ and $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$ under H_g. To simplify notation, we let $g(\alpha) = g(y \mid x; \alpha)$ and $h(\beta) = h(y \mid x; \beta)$.

Property 22.1: *Under the hypothesis H_g, the random vectors $\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))$ and $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$ are asymptotically normally distributed with zero means and respective variance covariance matrices*

$$\begin{aligned}\mathbf{V} &= \mathbf{K}_h^{-1}(\mathbf{C}_{hh} - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh})\mathbf{K}_h^{-1},\\ \mathbf{V}^* &= \mathbf{K}_h^{-1}(\mathbf{C}_{hh}^* - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh})\mathbf{K}_h^{-1},\end{aligned}$$

where

$$\begin{aligned}\mathbf{C}_{gg} &= E_x E_{\alpha_0}\left(\frac{\partial \log g(\alpha_0)}{\partial \alpha}\frac{\partial \log g(\alpha_0)}{\partial \alpha'}\right)\\ &= E_x E_{\alpha_0}\left(-\frac{\partial^2 \log g(\alpha_0)}{\partial \alpha \partial \alpha'}\right),\\ \mathbf{C}_{gh} &= E_x E_{\alpha_0}\left(\frac{\partial \log g(\alpha_0)}{\partial \alpha}\frac{\partial \log h(b(\alpha_0))}{\partial \beta'}\right),\\ \mathbf{C}_{hg} &= \mathbf{C}_{gh}',\\ \mathbf{C}_{hh} &= E_x E_{\alpha_0}\left(\frac{\partial \log h(b(\alpha_0))}{\partial \beta}\frac{\partial \log h(b(\alpha_0))}{\partial \beta'}\right),\\ \mathbf{K}_h &= -E_x E_{\alpha_0}\left(\frac{\partial^2 \log h(b(\alpha_0))}{\partial \beta \partial \beta'}\right),\\ \mathbf{C}_{hh}^* &= E_x E_{\alpha_0}\left(\left(\frac{\partial \log h(b(\alpha_0))}{\partial \beta} - E_{\alpha_0}\left(\frac{\partial \log h(b(\alpha_0))}{\partial \beta}\right)\right)\right.\\ &\quad \left.\times\left(\frac{\partial \log h(b(\alpha_0))}{\partial \beta} - E_{\alpha_0}\left(\frac{\partial \log h(b(\alpha_0))}{\partial \beta}\right)\right)'\right),\end{aligned}$$

and E_{α_0} denotes the expectation with respect to the probability distribution with density $g(y \mid x; \alpha_0)$.

PROOF: See Appendix 1 to this chapter. □

Let $\mathbf{V}^-$ and $\mathbf{V}^{*-}$ denote some generalized inverses of $\mathbf{V}$ and $\mathbf{V}^*$. Let $\hat{\mathbf{V}}^-$ and $\hat{\mathbf{V}}^{*-}$ be some consistent estimators of these generalized inverses. Let d and d^* denote the ranks of $\mathbf{V}$ and $\mathbf{V}^*$, respectively. Then it follows from Property 22.1 that the two test statistics

$$\xi_n^{W_1} = n(\hat{\beta}_n - b(\hat{\alpha}_n))'\hat{\mathbf{V}}^{-1}(\hat{\beta}_n - b(\hat{\alpha}_n)) \tag{22.13}$$

and

$$\xi_n^{W_2} = n(\hat{\beta}_n - b_n(\hat{\alpha}_n))'\hat{\mathbf{V}}^{*-}(\hat{\beta}_n - b_n(\hat{\alpha}_n)) \tag{22.14}$$

are asymptotically distributed as $\chi^2(d)$ and $\chi^2(d^*)$, respectively, under H_g. Thus we can propose two testing procedures with asymptotic level α based on the critical regions

$$\{\xi_n^{W_1} \geq \chi^2_{1-\alpha}(d)\} \quad \text{and} \quad \{\xi_n^{W_2} \geq \chi^2_{1-\alpha}(d^*)\}.$$

Remark 22.2: It is important to note that the asymptotic pseudo true value $b(\alpha_0)$ depends on the true joint distribution of (Y_i, X_i). In general, however, the marginal distribution of the exogenous variables X_i is left unspecified. This makes the computation of $b(.)$ impossible. Therefore the test based on $\xi_n^{W_1}$ is no longer feasible.

Remark 22.3: Consistent estimators of the matrices appearing in the expressions for $\mathbf{V}$ and $\mathbf{V}^*$ are obtained by replacing α_0 by $\hat{\alpha}_n$, $b(\alpha_0)$ by either $\hat{\beta}_n$ or $b_n(\hat{\alpha}_n)$, and the various expectations by their empirical means.

Remark 22.4: The difference between the asymptotic variance covariance matrices is

$$\begin{aligned}\mathbf{V} - \mathbf{V}^* &= \mathbf{K}_h^{-1}(\mathbf{C}_{hh} - \mathbf{C}_{hh}^*)\mathbf{K}_h^{-1} \\ &= \mathbf{K}_h^{-1} E_x \left(E_{\alpha_0}\left(\frac{\partial \log h(b(\alpha_0))}{\partial \beta}\right) E_{\alpha_0}\left(\frac{\partial \log h(b(\alpha_0))}{\partial \beta}\right)' \right) \mathbf{K}_h^{-1}.\end{aligned}$$

This difference is positive semidefinite. This implies that d is larger or equal to d^*.

Remark 22.5: The preceding results are consistent with those obtained when the models are nested. The latter case corresponds to hypotheses defined by

$$\begin{aligned} g(y_i \mid x_i; \alpha) &= k(y_i \mid x_i; 0, \alpha) \\ h(y_i \mid x_i; \beta) &= k(y_i \mid x_i; \gamma, \alpha), \text{ with } \beta' = (\gamma', \alpha'). \end{aligned}$$

Then we have

$$b(\alpha) = b_n(\alpha) = \begin{pmatrix} 0 \\ \alpha \end{pmatrix}$$

and

$$\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n)) = \sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n)) = \sqrt{n} \begin{pmatrix} \hat{\gamma}_n \\ \hat{\alpha}_n - \hat{\alpha}_n^0 \end{pmatrix},$$

where $\hat{\beta}_n = (\hat{\gamma}_n' \hat{\alpha}_n')'$ is the unconstrained maximum likelihood estimator of β and $\hat{\alpha}_n^0$ is the maximum likelihood estimator of α constrained by $\gamma = 0$. Hence in the case of nested models, the tests based on $\xi_n^{W_1}$ or $\xi_n^{W_2}$ coincide with the classical Hausman–Wald type test (see Chapter 7). This explains the name of generalized Wald tests given to these tests.

22.2.3 Generalized Score Tests

Some natural generalizations of the classical score test consist of evaluating the discrepancy from zero of either one of the two quantities

$$\hat{\lambda}_n^1 = \sum_{i=1}^n \frac{\partial \log h(Y_i \mid x_i; b(\hat{\alpha}_n))}{\partial \beta}$$

and

$$\hat{\lambda}_n^2 = \sum_{i=1}^n \frac{\partial \log h(Y_i \mid x_i; b_n(\hat{\alpha}_n))}{\partial \beta}.$$

Again, in the case of nested models these two statistics are identical to the score vector evaluated at the constrained maximum likelihood estimator. See Remark 22.5.

The asymptotic distributions of the statistics $\hat{\lambda}_n^1/\sqrt{n}$ and $\hat{\lambda}_n^2/\sqrt{n}$ are given in the next property.

Property 22.2: *Under H_g, the random vectors $\hat{\lambda}_n^1/\sqrt{n}$ and $\hat{\lambda}_n^2/\sqrt{n}$ are asymptotically normally distributed with zero means and respective variance covariance matrices*

$$\mathbf{W} = \mathbf{C}_{hh} - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh}$$

and

$$\mathbf{W}^* = \mathbf{C}_{hh}^* - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh}.$$

PROOF: See Appendix 1 to this chapter. □

As a matter of fact, the proof of Property 22.2 also shows that the random vectors $(1/\sqrt{n})\hat{\lambda}_n^1$ and $(1/\sqrt{n})\hat{\lambda}_n^2$ are asymptotically equivalent to $\mathbf{K}_h\sqrt{n}(b(\hat{\alpha}_n) - \hat{\beta}_n)$ and $\mathbf{K}_h\sqrt{n}(b_n(\hat{\alpha}_n) - \hat{\beta}_n)$, respectively, under H_g. This implies that the tests based on $\hat{\lambda}_n^1$ and $\hat{\lambda}_n^1$ are asymptotically equivalent to the generalized Wald tests based on the statistics $\xi_n^{W_1}$ and $\xi_n^{W_2}$, respectively.

From Property 22.2 it follows that the generalized score tests at asymptotic level α are defined by the following critical regions

$$\{\xi_n^{S_1} \geq \chi^2_{1-\alpha}(d)\} \quad \text{and} \quad \{\xi_n^{S_2} \geq \chi^2_{1-\alpha}(d^*)\},$$

where

$$\xi_n^{S_1} = \frac{1}{n}\hat{\lambda}_n^{1\prime}\hat{\mathbf{W}}^{-}\hat{\lambda}_n^1 \tag{22.15}$$

and

$$\xi_n^{S_2} = \frac{1}{n}\hat{\lambda}_n^{2\prime}\hat{\mathbf{W}}^{*-}\hat{\lambda}_n^2. \tag{22.16}$$

22.2.4 Selection of Regressors in Linear Models

a) The Pseudo True Values in Finite Samples

An important step in econometric modelling is the choice of the exogenous variables. In this section we study such a problem within the framework of the linear model.

The hypothesis H_g is defined by

$$Y = \mathbf{X}_2\gamma + u, \quad u \sim N(0, \sigma^2\mathbf{I}),$$

while the hypothesis H_h is defined by

$$Y = \mathbf{X}_1\delta + v, \quad v \sim N(0, \tau^2\mathbf{I}).$$

The matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are of dimension $(n \times k_1)$ and $(n \times k_2)$, and their ranks are equal to k_1 and k_2, respectively. The parameters α and β of the general theory presented above are here

$$\alpha = \begin{pmatrix} \gamma \\ \sigma^2 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \delta \\ \tau^2 \end{pmatrix}. \tag{22.17}$$

We consider the generalized Wald test procedure associated with the statistic $\xi_n^{W_2}$, i.e., the procedure based on the discrepancy between the maximum likelihood estimator of $\beta = (\delta', \tau^2)'$ and the estimator of the pseudo true value in finite samples, denoted $b_n(\hat{\alpha}_n)$. We have

$$\hat{\beta}_n = \begin{pmatrix} \hat{\delta}_n \\ \hat{\tau}_n^2 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'Y \\ \\ \frac{1}{n}\|(\mathbf{I} - \mathbf{M}_1)Y\|^2 \end{pmatrix}, \tag{22.18}$$

where $\mathbf{M}_1$ denotes the projection matrix $\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$.

Using equation (22.2), we obtain

$$b_n(\hat{\alpha}_n) = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\gamma}_n \\ \frac{1}{n}\|(\mathbf{I}-\mathbf{M}_2)Y\|^2 + \frac{1}{n}\|(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2\hat{\gamma}_n\|^2 \end{pmatrix}$$
$$= \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2 Y \\ \frac{1}{n}\|(\mathbf{I}-\mathbf{M}_2)Y\|^2 + \frac{1}{n}\|(\mathbf{I}-\mathbf{M}_1)\mathbf{M}_2 Y\|^2 \end{pmatrix},$$

where

$$\hat{\gamma}_n = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y \text{ and } \mathbf{M}_2 = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'.$$

b) Difference Between Estimators

From the preceding two equations we obtain

$$\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$$
$$= \sqrt{n}\begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(Y - \mathbf{X}_2\hat{\gamma}_n) \\ \frac{1}{n}(\|(\mathbf{I}-\mathbf{M}_1)Y\|^2 - \|(\mathbf{I}-\mathbf{M}_2)Y\|^2 - \|(\mathbf{I}-\mathbf{M}_1)\mathbf{M}_2 Y\|^2) \end{pmatrix}. \tag{22.19}$$

First, we shall show that the second set of components of $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$ is asymptotically a linear combination of the first k_1 components, under H_g. We have

$$\frac{1}{\sqrt{n}}(\|(\mathbf{I}-\mathbf{M}_1)Y\|^2 - \|(\mathbf{I}-\mathbf{M}_2)Y\|^2 - \|(\mathbf{I}-\mathbf{M}_1)\mathbf{M}_2 Y\|^2)$$
$$= \frac{1}{\sqrt{n}}(\|\mathbf{M}_1\mathbf{M}_2 Y\|^2 - \|\mathbf{M}_1 Y\|^2)$$
$$= \frac{1}{\sqrt{n}}(Y'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'Y$$
$$-Y'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'Y). \tag{22.20}$$

Now, under H_g, we have $Y = \mathbf{X}_2\gamma + u$. Hence it follows that another expression for the second set of components of $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$ is

$$\frac{1}{\sqrt{n}}(u'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'u$$
$$+ \quad 2\gamma'\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'u$$
$$- \quad u'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'u - 2\gamma'\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'u).$$

It is easy to see that the first and third terms converge to zero. Thus we are left with

$$\begin{aligned} & -\frac{2}{\sqrt{n}}\gamma'\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{I}-\mathbf{M}_2)u \\ &= -\frac{2}{\sqrt{n}}\gamma'\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(Y-\mathbf{X}_2\hat{\gamma}_n) \\ &= -2\gamma'\frac{1}{n}\mathbf{X}_2'\mathbf{X}_1\sqrt{n}(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(Y-\mathbf{X}_2\hat{\gamma}_n). \end{aligned} \tag{22.21}$$

The desired asymptotic equivalence follows.

From this asymptotic equivalence it follows that the generalized Wald test can be based upon the first k_1 components of $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$ only, or equivalently upon

$$\frac{1}{\sqrt{n}}\mathbf{X}_1'(Y-\mathbf{X}_2\hat{\gamma}_n) = \frac{1}{\sqrt{n}}\mathbf{X}_1'\hat{u}, \tag{22.22}$$

where $\hat{u}$ denotes the vector of estimated residuals in the regression defining H_g.

c) An Interpretation of the Test Statistic

The preceding generalized Wald test statistic is the scalar product of the vector of estimated residuals and the omitted variables $\mathbf{X}_1$. Thus, from equation (17.38), it is the score test statistic for testing the null hypothesis $H_0 : \delta = 0$ in the augmented linear regression model

$$Y = \mathbf{X}_2\gamma + \mathbf{X}_1\delta + w.$$

As a result, the generalized Wald test is asymptotically equivalent to a Fisher test in the above nesting model. Lastly, we note that if some columns of $\mathbf{X}_1$ belong to the subspace generated by the columns of $\mathbf{X}_2$, then we can suppress these columns when performing the test.

22.2.5 Choice Among Qualitative Variable Models

a) Statement of the Problem

We consider a qualitative variable that can take K values. We let Y_i denote the vector of size K whose k-th component Y_{ik}, $k = 1, \ldots, K$, is a random variable equal to 1 if the i-th observation falls in the k-th cell, and equal to zero otherwise.

We consider now two competing models H_g and H_h defined by

$$\begin{cases} H_g & : \quad \{P(Y_{ik}=1) = \tilde{g}_k(x_i;\alpha)\}, \\ H_h & : \quad \{P(Y_{ik}=1) = \tilde{h}_k(x_i;\beta)\}, \end{cases} \tag{22.23}$$

where the functions $\tilde{g}_k(x_i;\alpha)$ and $\tilde{h}_k(x_i;\beta)$ are such that $\tilde{g}_k(x_i;\alpha) > 0$, $\tilde{h}_k(x_i;\beta) > 0$, $\sum_{k=1}^K \tilde{g}_k(x_i;\alpha) = 1$ and $\sum_{k=1}^K \tilde{h}_k(x_i;\beta) = 1$.

The preceding formulation includes various situations. For instance, consider a quantitative variable Y_i^* that satisfies the linear model $Y_i^* = x_i b + u_i$, where the errors u_i are independently and identically distributed with common probability distribution P. Suppose that we only observe the qualitative variable $Y_i = \mathbb{1}_{Y_i^* \geq 0}$. Then, the preceding formulation allows to study not only the problem of selecting the explanatory variables x but also the problem of choosing the probability distribution of the errors u_i.

Assuming that the observations are independent, the log-likelihood corresponding to the two models are

$$L_n^g(\alpha) = \sum_{i=1}^n \sum_{k=1}^K y_{ik} \log \tilde{g}_k(x_i;\alpha) = \sum_{i=1}^n \log g(y_i \mid x_i;\alpha) \tag{22.24}$$

and

$$L_n^h(\beta) = \sum_{i=1}^n \sum_{k=1}^K y_{ik} \log \tilde{h}_k(x_i;\beta) = \sum_{i=1}^n \log h(y_i \mid x_i;\beta). \tag{22.25}$$

b) The Estimated Pseudo True Values

The maximization of the log-likelihood functions provides the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$. These estimators depend on the observations. To compute $b_n(\hat{\alpha}_n)$, we minimize

$$E_{\hat{\alpha}_n} L_n^h(\beta) = \sum_{i=1}^n \sum_{k=1}^K \tilde{g}_k(x_i;\hat{\alpha}_n) \log \tilde{h}_k(x_i;\beta). \tag{22.26}$$

Comparing equation (22.26) to equation (22.25), we see that $b_n(\hat{\alpha}_n)$ is obtained by maximizing the log-likelihood function $L_n^h(\beta)$, where the observations y_{ik} have been replaced by their optimal predictions $\tilde{g}_k(x_i;\hat{\alpha}_n)$ computed under H_g. Thus we have

$$b_n(\hat{\alpha}_n) = \alpha\left(\tilde{g}_k(x_i;\hat{\alpha}_n)\right).$$

c) Generalized Score Test

For instance, consider applying the generalized score test based on $\xi_n^{S_2}$. It is necessary to have some expressions for the matrices $\mathbf{C}_{hh}^*$, $\mathbf{C}_{hg}$, and $\mathbf{C}_{gg}$. These expressions can be obtained from the form of the score vectors, namely

$$\begin{aligned}\frac{\partial \log h(y \mid x; b(\alpha))}{\partial \beta} - E_\alpha \frac{\partial \log h(y \mid x; b(\alpha))}{\partial \beta} \\ = \sum_{k=1}^{K} \frac{y_k - \tilde{g}_k(x;\alpha)}{\tilde{h}(x; b(\alpha))} \frac{\partial \tilde{h}_k(x; b(\alpha))}{\partial \beta}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \log g(y \mid x; \alpha)}{\partial \alpha} &= \sum_{k=1}^{K} \frac{y_k}{\tilde{g}_k(x;\alpha)} \frac{\partial \tilde{g}_k(x;\alpha)}{\partial \alpha} \\ &= \sum_{k=1}^{K} \frac{y_k - \tilde{g}_k(x;\alpha)}{\tilde{g}_k(x;\alpha)} \frac{\partial \tilde{g}_k(x;\alpha)}{\partial \alpha}.\end{aligned}$$

For instance, we have

$$\begin{aligned}\mathbf{C}_{hh}^* = E_x \Bigg(&\frac{\partial \tilde{h}'(x; b(\alpha_0))}{\partial \beta} \operatorname{diag}\,(1/\tilde{h}(x; b(\alpha_0))) \\ &\times (\operatorname{diag}\,\tilde{g}(x;\alpha_0) - \tilde{g}(x;\alpha_0)\tilde{g}'(x;\alpha_0)) \\ &\times \operatorname{diag}\,(1/\tilde{h}(x; b(\alpha_0))) \frac{\partial \tilde{h}(x; b(\alpha_0))}{\partial \beta'}\Bigg),\end{aligned}$$

where $\tilde{h}$ is the vector whose k-th component is $\tilde{h}_k$, and $\partial \tilde{h}'/\partial \beta$ is the matrix whose k-th column vector is $\partial \tilde{h}_k/\partial \beta$. Similar expressions can be obtained for the other two matrices $\mathbf{C}_{hg}$ and $\mathbf{C}_{gg}$.

Now, these matrices can be estimated consistently by replacing E_x by an empirical mean, α_0 by $\hat{\alpha}_n$, and $b(\alpha_0)$ by $b_n(\hat{\alpha}_n)$. Thus, for instance, a consistent estimator of the matrix $\mathbf{C}_{hh}^*$ is

$$\begin{aligned}\hat{\mathbf{C}}_{hh}^* = \frac{1}{n} \sum_{i=1}^{n} \Bigg(&\frac{\partial \tilde{h}'(x_i; b_n(\hat{\alpha}_n))}{\partial \beta} \operatorname{diag}\,(1/\tilde{h}(x_i; b_n(\hat{\alpha}_n))) \\ &\times (\operatorname{diag}(\tilde{g}(x_i;\hat{\alpha}_n)) - \tilde{g}(x_i;\hat{\alpha}_n)\tilde{g}'(x_i;\hat{\alpha}_n)) \\ &\times \operatorname{diag}(1/\tilde{h}(x_i; b_n(\hat{\alpha}_n))) \frac{\partial \tilde{h}(x_i; b_n(\hat{\alpha}_n))}{\partial \beta'}\Bigg).\end{aligned}$$

Then we can compute the test statistic $\xi_n^{S_2}$ and apply the generalized score test procedure. See Exercise 22.3 for an illustration.

22.2.6 Cox Testing Procedure

We consider again the preceding framework. Specifically, we consider two competing hypotheses H_g and H_h defined by the following two families of conditional probability densities

$$\begin{aligned} H_g &: \{g(y_i \mid x_i; \alpha), \alpha \in A \subset \mathbb{R}^G\}, \\ H_h &: \{h(y_i \mid x_i; \beta), \beta \in B \subset \mathbb{R}^H\}. \end{aligned}$$

We want to test H_g against H_h (say). In the preceding sections we have generalized the Wald test and the score test. Then it is natural to ask whether the third classical asymptotic test, namely the test based on the likelihood ratio, can be generalized in such a testing situation. We let

$$L_n^g(\alpha) = \sum_{i=1}^n \log g(y_i \mid x_i; \alpha)$$

and

$$L_n^h(\beta) = \sum_{i=1}^n \log h(y_i \mid x_i; \beta)$$

denote the log-likelihood functions under H_g and H_h, respectively. Then the statistic generalizing the likelihood ratio test statistic would be

$$\begin{aligned} \xi_n^M &= 2(L_n^h(\hat{\beta}_n) - L_n^g(\hat{\alpha}_n)) \\ &= 2\sum_{i=1}^n (\log h(y_i \mid x_i; \hat{\beta}_n) - \log g(y_i \mid x_i; \hat{\alpha}_n)). \end{aligned}$$

Under H_g, however, as n increases to infinity, the statistic $(1/n)\xi_n^M$ converges to

$$2E_x E_{\alpha_0}(\log h(y \mid x; b(\alpha_0)) - \log g(y \mid x; \alpha_0)),$$

which is different from zero, in general. Hence $|\xi_n^M|$ diverges to $+\infty$. Therefore ξ_n^M cannot be asymptotically equivalent to the generalized Wald and score test statistics, which are asymptotically distributed as χ^2.

Following Cox, this remark allows us to propose a test statistic converging to zero under H_g as n increases to infinity, namely

$$C_n = \frac{1}{n}(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{\beta}_n)) - H(\hat{\alpha}_n), \tag{22.27}$$

where

$$H(\alpha) = E_x E_\alpha(\log g(Y \mid X;\alpha) - \log h(Y \mid X; b(\alpha))). \tag{22.28}$$

In addition, to obtain a statistic that has asymptotically a nonsingular distribution under H_g, we shall consider the statistic $\sqrt{n}C_n$. We have the following property.

Property 22.3: *Under H_g, the test statistic $\sqrt{n}C_n$ is asymptotically normally distributed with zero mean and variance*

$$\begin{aligned}\omega_0^2 \;=\; & V\left(\log \frac{g(Y \mid X;\alpha_0)}{h(Y \mid X; b(\alpha_0))}\right) \\ & -\operatorname{cov}\left(\log \frac{g(Y \mid X;\alpha_0)}{h(Y \mid X; b(\alpha_0))}, \frac{\partial \log g(Y \mid X;\alpha_0)}{\partial \alpha}\right) \\ & \times \left(V\left(\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial \alpha}\right)\right)^{-1} \\ & \times \operatorname{cov}\left(\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial \alpha}, \log \frac{g(Y \mid X;\alpha_0)}{h(Y \mid X; b(\alpha_0))}\right),\end{aligned}$$

where α_0 denotes the true parameter value for α, and the various variances and covariances are evaluated with respect to the joint probability distribution of (Y, X), which is defined by the conditional probability distribution of Y given X with density $g(y \mid x;\alpha_0)$ and by the marginal probability distribution of X.

PROOF: See Appendix 2 to this chapter. □

Property 22.3 allows us to define the Cox testing procedure.

Property 22.4: Cox Test. *If ω_0^2 is nonzero and if $\hat{\omega}_n^2$ denotes a consistent estimator of ω_0^2 under H_g, then the test with critical region*

$$W_n = \left\{\frac{\sqrt{n}C_n}{\hat{\omega}_n} < u_\alpha\right\},$$

where u_α denotes the α-th order quantile of the $N(0,1)$ distribution, has asymptotic level equal to α.

Moreover, this test is consistent against the probability distributions in H_h that do not belong to H_g.

PROOF:

(i) Property 22.3 implies at once that the asymptotic distribution of $\sqrt{n}C_n/\hat{\omega}_n$ is the $N(0,1)$ distribution under H_g if ω_0^2 is nonzero. It follows that the probability of W_n converges to α for every probability distribution belonging to H_g, as desired.

(ii) Suppose now that the true probability distribution is a probability distribution that belongs to H_h and hence that corresponds to some parameter value β_0, but that does not belong to H_g. Under such a condition, the estimator $\hat{\alpha}_n$ converges to $a(\beta_0)$ while the estimator $\hat{\beta}_n$ converges to β_0. Hence

$$\frac{1}{n}(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{\beta}_n)) \to E_x E_{\beta_0} \log \frac{g(Y \mid X; a(\beta_0))}{h(Y \mid X; \beta_0)}$$

while

$$H(\hat{\alpha}_n) \to E_x E_{a(\beta_0)} \log \frac{g(Y \mid X; a(\beta_0))}{h(Y \mid X; b(a(\beta_0)))}.$$

It follows that C_n converges to

$$C = E_x \left(E_{\beta_0} \log \frac{g(Y \mid X; a(\beta_0))}{h(Y \mid X; \beta_0)} - E_{a(\beta_0)} \log \frac{g(Y \mid X; a(\beta_0))}{h(Y \mid X; b(a(\beta_0)))} \right),$$

where E_{β_0} and $E_{a(\beta_0)}$ indicate that the expectations are taken with respect to the conditional probability densities $h(y \mid x; \beta_0)$ and $g(y \mid x; a(\beta_0))$, respectively.

Then Kullback inequality implies that each of the two terms in this limit is negative or zero. Moreover, if $h(y \mid x; \beta_0)$ does not belong to H_g, as assumed, then the first term is strictly negative. Hence C is strictly negative. It follows that $\sqrt{n}C_n$ converges to $-\infty$. Therefore the power of the test with critical region W_n converges to one, as desired. □

Applying the Cox test requires a consistent estimator $\hat{\omega}_n^2$ of ω_0^2. For instance, such an estimator can be obtained by replacing in the expression for ω_0^2: α_0 by $\hat{\alpha}_n$, $b(\alpha_0)$ by $\hat{\beta}_n$ or $b_n(\hat{\alpha}_n)$, and the mathematical expectations by the appropriate empirical means.

Example 22.1: The Linear Model

(i) THE COX TEST STATISTIC

We consider the same hypotheses as in Section 22.2.4, namely

$$\begin{aligned} H_g &: Y = \mathbf{X}_2\gamma + u, \quad u \sim N(0, \sigma^2\mathbf{I}), \\ H_h &: Y = \mathbf{X}_1\delta + v, \quad v \sim N(0, \tau^2\mathbf{I}). \end{aligned}$$

The parameters α and β are $\alpha = (\gamma', \sigma^2)'$ and $\beta = (\delta', \tau^2)'$.

We have

$$L_n^g(\alpha) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Y - \mathbf{X}_2\gamma\|^2,$$

and

$$L_n^h(\beta) = -\frac{n}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\|Y - \mathbf{X}_1\delta\|^2.$$

Hence

$$\begin{aligned} \frac{1}{n}(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{\beta}_n)) &= \frac{1}{n}\left(-\frac{n}{2}(2\pi\hat{\sigma}_n^2) - \frac{n}{2} + \frac{n}{2}\log(2\pi\hat{\tau}_n^2) + \frac{n}{2}\right) \\ &= \frac{1}{2}\log\frac{\hat{\tau}_n^2}{\hat{\sigma}_n^2}. \end{aligned}$$

Under H_g, as n increases to infinity, then $\hat{\sigma}_n^2$ converges to σ_0^2 while $\hat{\tau}_n^2$ converges to

$$t^2(\alpha_0) = \sigma_0^2 + \gamma_0'(E(x_{i2}'x_{i1}) - E(x_{i2}'x_{i1})(E(x_{i1}'x_{i1}))^{-1}E(x_{i1}'x_{i2}))\gamma_0$$

(see equation (22.4)). Hence we can take $H(\hat{\alpha}_n)$ to be

$$\frac{1}{2}\log\frac{\hat{\sigma}_n^2 + \hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I} - \mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n}{\hat{\sigma}_n^2}.$$

Therefore we have

$$\begin{aligned} C_n &= \frac{1}{n}(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{b}_n)) - H(\hat{\alpha}_n) \\ &= \frac{1}{2}\log\frac{\hat{\tau}_n^2}{\hat{\sigma}_n^2 + \hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I} - \mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n}. \end{aligned}$$

(ii) A STATISTIC THAT IS ASYMPTOTICALLY EQUIVALENT TO THE COX TEST STATISTIC

We have

$$\begin{aligned}\sqrt{n}C_n &= \frac{\sqrt{n}}{2}\log\frac{\hat{\tau}_n^2}{\hat{\sigma}_n^2+\hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n},\\ &\# \frac{\sqrt{n}}{2}\frac{\hat{\tau}_n^2-\hat{\sigma}_n^2-\hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n}{\hat{\sigma}_n^2+\hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n},\\ &\# \frac{\sqrt{n}}{2}\frac{\hat{\tau}_n^2-\hat{\sigma}_n^2-\hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n}{t^2(\alpha_0)}.\end{aligned}$$

On the other hand, from Section 22.2.4 and in particular equation (22.22) we know that

$$\begin{aligned}\sqrt{n}(\hat{\tau}_n^2-\hat{\sigma}_n^2-\hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n) &= -2\gamma'\mathbf{X}_2'\mathbf{M}_1\frac{Y-\mathbf{X}_2\hat{\gamma}_n}{\sqrt{n}}\\ &= -2\gamma'\mathbf{X}_2'\mathbf{M}_1(\mathbf{I}-\mathbf{M}_2)\frac{u}{\sqrt{n}}.\end{aligned}$$

Hence a statistic that is asymptotically equivalent to the Cox test statistic is

$$\xi_n = \frac{n(\hat{\tau}_n^2-\hat{\sigma}_n^2)-\hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n}{2\hat{\sigma}_n(\hat{\gamma}_n'\mathbf{X}_2'\mathbf{M}_1(\mathbf{I}-\mathbf{M}_2)\mathbf{M}_1\mathbf{X}_2\hat{\gamma}_n)^{1/2}}.$$

That is, from equation (22.20), the Cox test is asymptotically equivalent to a generalized Wald test using only the last component of $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$, i.e., the difference between the estimator $\hat{\tau}_n^2$ of the variance τ^2 of the error term under H_h and the estimator of the pseudo true value of τ^2 under H_g, namely

$$\hat{\sigma}_n^2+\hat{\gamma}_n'(\mathbf{X}_2'(\mathbf{I}-\mathbf{M}_1)\mathbf{X}_2)\hat{\gamma}_n/n.$$

22.2.7 Using a Nesting Model

We consider again the competing hypotheses

$$H_g : \{g(y_i \mid x_i;\alpha), \alpha \in A \subset I\!R^G\}$$

and

$$H_h : \{h(y_i \mid x_i;\beta), \beta \in B \subset I\!R^H\}.$$

In order to apply the classical testing procedures studied in Chapter 17, we can think of defining a nesting or *comprehensive* model containing

both H_g and H_h as special cases. Then we can test H_g (or H_h) against this comprehensive model.

Definition 22.4: *A comprehensive model for the hypotheses H_g and H_h is a family $\mathcal{M}$ of conditional probability densities containing the families associated with H_g and H_h.*

Given a comprehensive model $\mathcal{M}$, then in principle we can test the null hypothesis H_g (say) against the alternative hypothesis $\mathcal{M} - H_g$, which contains H_h by construction, using classical testing procedures. In this approach, there clearly exists some arbitrariness in the choice of the comprehensive model as well as in the choice of the null hypothesis.

We present below some commonly used comprehensive models that have been proposed.

a) Quandt and Atkinson Comprehensive Models

These comprehensive models consider families that are obtained by mixing the probability distributions in H_g and H_h. The mixture can be based on an arithmetic mean (Quandt) or a geometric mean (Atkinson). Specifically, the *Quandt comprehensive model* is defined by

$$\mathcal{M}=\{\lambda g(y_i \mid x_i;\alpha)+(1-\lambda)h(y_i \mid x_i;\beta), \lambda \in [0,1], \alpha \in A, \beta \in B\}. \tag{22.29}$$

The *Atkinson comprehensive model* is defined by

$$\mathcal{M}=\left\{\frac{g^{\lambda}(y_i \mid x_i;\alpha)h^{1-\lambda}(y_i \mid x_i;\beta)}{\int g^{\lambda}(y_i \mid x_i;\alpha)h^{1-\lambda}(y_i \mid x_i;\beta)d\nu(y_i)}, \lambda \in [0,1], \alpha \in A, \beta \in B\right\}. \tag{22.30}$$

The Atkinson comprehensive model requires the existence of the integral appearing in the denominator in equation (22.30). In both comprehensive models the hypothesis H_g is obtained by imposing $\lambda = 1$, while the hypothesis H_h is obtained by imposing $\lambda = 0$. Thus, in principle, we can test H_g or H_h within each comprehensive model by testing $\lambda = 1$ or $\lambda = 0$, respectively.

These comprehensive models raise, however, an identification problem since, for instance, β is not identified under H_g. This identification problem can be solved in some special cases as the next example illustrates.

Example 22.2: The hypothesis H_g is defined by

$$Y = \mathbf{X}b + u, \quad u \sim N(0, \sigma^2\mathbf{I}),$$

and the hypothesis H_h is defined by

$$Y = \mathbf{X}b + v, \quad v \sim N(0, \tau^2\mathbf{D}),$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal elements are z_i^2, $i = 1, \ldots, n$. Note that, because the same matrix of explanatory variables $\mathbf{X}$ enters in both models, we have already set the coefficient vector b to be identical across the two models so as to economize on the number of parameters in the comprehensive model.

The log-densities of the i-th observation under H_g and H_h are respectively

$$\log g(y_i \mid x_i; b, \sigma^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y_i - x_i b)^2$$

and

$$\log h(y_i \mid x_i, z_i; b, \tau^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log(\tau^2 z_i^2) - \frac{1}{2\tau^2 z_i^2}(y_i - x_i b)^2.$$

Thus the Atkinson comprehensive model is defined by log-densities of the form

$$-\frac{1}{2}\log 2\pi + \frac{1}{2}\log\left(\frac{\lambda}{\sigma^2} + \frac{1-\lambda}{\tau^2 z_i^2}\right) - \frac{1}{2}\left(\frac{\lambda}{\sigma^2} + \frac{1-\lambda}{\tau^2 z_i^2}\right)(y_i - x_i b)^2.$$

Hence the Atkinson comprehensive model corresponds to a family of normal distributions with means $x_i b$ and variances $\left((\lambda/\sigma^2) + (1-\lambda)/\tau^2 z_i^2\right)^{-1}$. Note that the latter variance is a harmonic mean of the original variances. Because the variance is a function of $(\lambda, \tau^2, \sigma^2)$ through λ/σ^2, it is clear that any given value of this variance can be obtained from an infinity of values of the triple $(\lambda, \tau^2, \sigma^2)$. Therefore the Atkinson comprehensive model is not identified.

Nevertheless the identification problem can be solved by writing

$$\left(\frac{\lambda}{\sigma^2} + \frac{1-\lambda}{\tau^2 z_i^2}\right)^{-1} = \frac{1}{\gamma + \delta/z_i^2}, \quad \text{with} \quad \gamma = \frac{\lambda}{\sigma^2}, \delta = \frac{1-\lambda}{\tau^2}.$$

Then the comprehensive model is identified in (b, γ, δ). The hypothesis H_g is obtained by imposing $\delta = 0$, while the hypothesis H_h is obtained by imposing $\gamma = 0$. However, because the parameters δ and γ are nonnegative by construction, one faces a testing problem on the boundaries of the parameter space of the type studied in Chapter 21.

b) The Box–Cox Comprehensive Model

The Box–Cox comprehensive model is obtained by applying the so-called the Box–Cox transformation, which is defined by

$$f_\lambda(z) = \frac{z^\lambda - 1}{\lambda}, \text{ for } \lambda \neq 0 \text{ and } f_0(z) = \log z, \tag{22.31}$$

where z is a strictly positive variable. Note that $f_0(z)$ is the limit of $f_\lambda(z)$ as λ decreases to zero. For we have

$$\frac{z^\lambda - 1}{\lambda} = \frac{\exp(\lambda \log z) - 1}{\lambda} \# \log z.$$

On the other hand, when $\lambda = 1$, we obtain $z - 1$, i.e., we obtain the identity function up to some addtitive term. Hence the Box–Cox transformation can be used to make a choice between a specification in the logarithm of a strictly positive variable and a specification in the level of that variable.

For instance, the regression model

$$Y_i = b_1 + b_2 f_\lambda(x_i) + u_i, \quad i = 1, \ldots, n, \tag{22.32}$$

contains, as special cases, the model $Y_i = b_1^* + b_2 x_i + u_i$, where $b_1^* = b_1 - b_2$, by setting $\lambda = 1$, and the model $Y_i = b_1 + b_2 \log x_i + u_i$, by setting $\lambda = 0$. Thus we can apply the classical procedures for testing the hypotheses $\lambda = 0$ or $\lambda = 1$ so as to choose between a linear specification and a semi-logarithmic specification.

The Box–Cox transformation can also be used to choose between a specification in Y_i or a specification in $\log Y_i$ in a regression model. Specifically, the comprehensive model is

$$f_\lambda(Y_i) = x_i b + u_i. \tag{22.33}$$

Note that the variable Y_i must be strictly positive. Thus, in principle, this excludes a model of the form

$$Y_i = x_i b + u_i, \quad u_i \sim N(0, \sigma^2),$$

since Y_i can be negative in such a model. However, if the mean $x_i b$ is strongly positive and large relative to the standard error σ, then Y_i will be negative with a very small probability, in which case, the normality of the errors can be viewed as an approximation.

c) The Davidson–McKinnon Comprehensive Model

(i) LINEAR REGRESSIONS

We consider the following two linear models

$$H_g : Y_i = x_i b + u_i,\ E(u_i \mid x_i) = 0,\ V(u_i \mid x_i) = \sigma^2$$

and

$$H_h : Y_i = z_i c + v_i,\ E(v_i \mid z_i) = 0,\ V(v_i \mid z_i) = \tau^2.$$

Then Davidson and McKinnon introduce the comprehensive model

$$\mathcal{M} : Y_i = (1-\lambda)x_i b + \lambda z_i c + w_i. \tag{22.34}$$

However, if one wants to test H_g in $\mathcal{M}$, i.e., $\lambda = 0$, then the parameter vector c becomes unidentified under H_g. To solve this identification problem, Davidson and McKinnon consider the *pseudo model*

$$Y_i = x_i b + \lambda z_i \hat{c}_n + \epsilon_i, \tag{22.35}$$

where $\hat{c}_n$ is the ordinary least squares estimator of c under the hypothesis H_h. Such a method is justified by the next property.

Property 22.5: *Under H_g, the Student statistic associated with λ in the ordinary least squares estimation of the pseudo model (22.35) is asymptotically $N(0,1)$ distributed if x_i and z_i are such that $Ex_i'z_i \neq 0$.*

PROOF: The ordinary least squares estimator of λ in the pseudo model (22.35) is

$$\hat{\lambda}_n = (\hat{c}_n' \mathbf{Z}'(\mathbf{I} - \mathbf{M}_x)\mathbf{Z}\hat{c}_n)^{-1}\hat{c}_n \mathbf{Z}'(\mathbf{I} - \mathbf{M}_x)Y, \tag{22.36}$$

where $\mathbf{Z}$ is the matrix whose i-th row is z_i, $\mathbf{X}$ is the matrix whose i-th row is x_i, and $\mathbf{M}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the matrix of the orthogonal projection on the subspace generated by the columns of $\mathbf{X}$.

The estimator of the variance of $\hat{\lambda}_n$, which is commonly computed in ordinary least squares estimation is

$$\hat{V}(\hat{\lambda}_n) = \hat{\sigma}^2(\hat{c}_n' \mathbf{Z}'(\mathbf{I} - \mathbf{M}_x)\mathbf{Z}\hat{c}_n)^{-1},$$

where $\hat{\sigma}^2$ denotes the usual estimator of the variance of the errors.

Therefore the Student test statistic associated with λ is

$$T_\lambda = \frac{\hat{c}_n' \mathbf{Z}'(\mathbf{I} - \mathbf{M}_x)Y}{\hat{\sigma}(\hat{c}_n' \mathbf{Z}'(\mathbf{I} - \mathbf{M}_x)\mathbf{Z}\hat{c}_n)^{1/2}}. \tag{22.37}$$

But, under the hypothesis H_g, we have $Y = \mathbf{X}b_0 + u$. Thus T_λ can be written as

$$T_\lambda = \frac{\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)u}{\hat{\sigma}(\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)\mathbf{Z}\hat{c}_n)^{1/2}} = \frac{\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)u/\sqrt{n}}{\hat{\sigma}\left(\frac{1}{n}\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)\mathbf{Z}\hat{c}_n\right)^{1/2}}.$$

When n increases to infinity, then $\hat{c}_n$ converges to the pseudo true value for c, i.e., to

$$c_0^* = (E(z_i'z_i))^{-1}E(z_i'x_i)b_0,$$

while $\hat{\sigma}^2$ converges to the true value σ_0^2 of σ^2. Hence the denominator of T_λ converges to

$$\sigma_0(c_0^{*\prime}(E(z_i'z_i) - E(z_i'x_i)(E(x_i'x_i)^{-1}E(x_i'z_i))c_0^*)^{1/2},$$

while the numerator converges in distribution to

$$N(0, \sigma_0^2 c_0^{*\prime}(E(z_i'z_i) - E(z_i'x_i)E(x_i'x_i)^{-1}E(x_i'z_i))c_0^*).$$

The desired result follows.

If $E(x_i'z_i) = 0$, then $c_0^* = 0$. Hence the above proof is no longer valid. See Exercise 22.5. □

Remark 22.6: Under the alternative hypothesis H_h, we have

$$\begin{aligned} T_\lambda &= \frac{\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)\mathbf{Z}c_0 + \hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)v}{\hat{\sigma}(\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)\mathbf{Z}\hat{c}_n)^{1/2}} \\ &= \frac{\frac{1}{\sqrt{n}}\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)\mathbf{Z}c_0 + \frac{1}{\sqrt{n}}\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)v}{\hat{\sigma}\left(\frac{1}{n}\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)\mathbf{Z}\hat{c}_n'\right)^{1/2}}. \end{aligned}$$

The first term in the numerator can be written as

$$\begin{aligned} \sqrt{n}\frac{\hat{c}_n'\mathbf{Z}'(\mathbf{I}-\mathbf{M}_x)\mathbf{Z}c_0}{n} &\# \sqrt{n}c_0'(E(z_i'z_i) - E(z_i'x_i)(E(x_i'x_i))^{-1}E(x_i'z_i))c_0 \\ &= \sqrt{n}A, \end{aligned}$$

by definition. If z_ic_0 is not a linear combination of the components of x_i, then the A term is strictly positive. Therefore $\sqrt{n}A$ diverges to $+\infty$. Hence, because the second term is negligible relative to the first term, the statistic T_λ also diverges to $+\infty$. It follows that the one-sided test with critical region $\{T_\lambda \geq u_{1-\alpha}\}$ is consistent against the alternative hypothesis H_h.

(ii) NONLINEAR REGRESSIONS: THE J TEST

The preceding approach can be generalized to the case of nonlinear regressions. Specifically, consider the competing nonlinear regression models

$$\begin{aligned} H_g &: Y_i = g(x_i; b) + u_i, \\ H_h &: Y_i = h(z_i; c) + v_i. \end{aligned}$$

The Davidson–McKinnon comprehensive model is

$$\mathcal{M} : Y_i = (1-\lambda)g(x_i; b) + \lambda h(z_i; c) + w_i, \tag{22.38}$$

which is replaced by the pseudo model

$$Y_i = (1-\lambda)g(x_i; b) + \lambda h(z_i; \hat{c}_n) + \epsilon_i, \tag{22.39}$$

where $\hat{c}_n$ is the nonlinear least squares estimator of c under H_h.

Then the joint nonlinear least squares estimation of λ and b in the pseudo model (22.39) provides a Student test statistic for λ. This statistic is asymptotically $N(0,1)$ distributed under H_g and can be used to construct a test. Such a test is called the *J test*, where J stands for joint.

(iii) NONLINEAR REGRESSIONS: THE P TEST

We can linearize the pseudo model (22.39) around $\lambda = 0$ and $b = \hat{b}_n$, where $\hat{b}_n$ is the nonlinear least squares estimator of b under H_g. Then we obtain a relation of the form

$$Y_i - g(x_i; \hat{b}_n) = \lambda(h(z_i; \hat{c}_n) - g(x_i; \hat{b}_n)) + \hat{G}_i\beta + \eta_i,$$

where $\hat{G}_i$ is the row vector of derivatives of $g(x_i; b)$ with respect to b evaluated at $\hat{b}_n$.

In the resulting pseudo model, which is linear in λ and β, the Student test statistic associated with λ is

$$\tilde{T}_\lambda = \frac{(\hat{h} - \hat{g})'(\mathbf{I} - \hat{\mathbf{M}}_G)(y - \hat{g})}{\hat{\sigma}((\hat{h} - \hat{g})'(\mathbf{I} - \hat{\mathbf{M}}_G)(\hat{h} - \hat{g}))^{1/2}}, \tag{22.40}$$

where $\hat{\mathbf{M}}_G$ is the orthogonal projection matrix onto the subspace generated by the columns of the matrix $\hat{\mathbf{G}}$ whose i-th row is $\hat{G}_i$, $\hat{h}$ is the vector whose components are $h(z_i; \hat{c}_n)$, and $\hat{g}$ is the vector whose components are $g(x_i; \hat{b}_n)$.

Under H_g, the statistic $\tilde{T}_\lambda$ has the same asymptotic distribution as

$$\frac{(h - g)'(\mathbf{I} - \mathbf{M}_G)u}{\sigma((h - g)'(\mathbf{I} - \mathbf{M}_G)(h - g))^{1/2}},$$

where the matrice without hats correspond to the previous matrices with hats and $\hat{b}_n$ and $\hat{c}_n$ are replaced by the true value b_0 and the pseudo true values c_0^*, respectively.

A reasoning analogous to the one given for the linear case shows that the asymptotic distribution of $\tilde{T}_\lambda$ is $N(0,1)$ under H_g. Similarly, it can be shown that the statistic $\tilde{T}_\lambda$ converges to $+\infty$, in general, under H_h. Thus, a consistent test can be proposed based on the critical region $\{T_\lambda \geq u_{1-\alpha}\}$. This test is called the *P test*, where P stands for the projection matrix $\mathbf{I}-\hat{\mathbf{M}}_G$ appearing in the definition of the test statistic.

22.3 Model Selection Criteria

As indicated in Section 22.2.1, the preceding methods may lead to some inconsistent outcomes when H_g and H_h are successively chosen to be the null hypothesis. In addition, this testing procedure is not well adapted to the problem of choosing among more than two competing models. These two reasons among others have led to the use of model selection criteria.

A model selection criterion is a real scalar function defined on the collection of competing models and on the sample space $\mathcal{Y}$. For a given observation $y \in \mathcal{Y}$, the procedure consists of retaining the model that maximizes (or minimizes) the chosen criterion. Hence a partitioning of the sample space $\mathcal{Y}$ is obtained, where each element of this partition corresponds to the choice of one model.

22.3.1 Linear Models

One of the criteria that is most commonly used in linear models is the coefficient of determination. We briefly recall its definition and its interpretation. Then we shall examine its consistency properties.

a) The Coefficient of Determination

Consider the linear model

$$Y = \mathbf{X}b + u, \quad E(u \mid \mathbf{X}) = 0, V(u \mid \mathbf{X}) = \sigma^2\mathbf{I}, \tag{22.41}$$

where the first column of the matrix $\mathbf{X}$ is the vector e whose components are all equal to 1.

Let $\hat{b}_n$ be the ordinary least squares estimator of b, $\hat{Y} = \mathbf{X}\hat{b}_n$ be the vector of predicted values, and $\hat{u} = Y - \hat{Y}$ be the vector of residuals. The analysis-of-variance equation is

$$\|Y - \bar{Y}e\|^2 = \|\hat{Y} - \bar{Y}e\|^2 + \|\hat{u}\|^2 \tag{22.42}$$

Then the *coefficient of determination* or squared *multiple correlation coefficient* measures the part of the variance that is explained by the model, namely

$$R^2 = \frac{\|\hat{Y} - \bar{Y}e\|^2}{\|Y - \bar{Y}e\|^2} = 1 - \frac{\|\hat{u}\|^2}{\|Y - \bar{Y}e\|^2}. \tag{22.43}$$

From the analysis-of-variance equation, it follows that the coefficient of determination is always between 0 and 1.

A choice among competing linear models based on R^2 consists of selecting the model that provides the "best" predictions over the observed sample. Thus, intuitively, such a model selection criterion is not directly related to the explanatory power of an econometric model.

When the errors are normally distributed, it is interesting to note that the Fisher statistic (see Section 16.2.3) corresponding to the hypothesis that all the components of b are equal to zero with the exception of the coefficient associated with the constant term is

$$F = \frac{\|\hat{Y} - \bar{Y}e\|^2/(K-1)}{\|\hat{u}\|^2/(n-K)} \tag{22.44}$$

and

$$F = \frac{n-K}{K-1}\left(\frac{1}{1-R^2} - 1\right) = \frac{n-K}{K-1}\,\frac{R^2}{1-R^2}. \tag{22.45}$$

Therefore the F statistic is an increasing function of the coefficient of determination R^2. Hence the critical region of the Fisher test is implicitly based on R^2.

b) The Adjusted Coefficient of Determination

The preceding model selection criterion displays, however, some undesirable properties. Specifically, it is easy to see that if another explanatory variable is added to the model (22.41), i.e., if another column is added to the matrix $\mathbf{X}$, then the sum of squared residuals $\|\hat{u}\|^2$ will decrease and, hence, the coefficient of determination R^2 will increase. This obvious remark implies that R^2 cannot be taken as a criterion for choosing

among competing linear models since, according to this criterion, the addition of any explanatory variable will always increase the "quality" of the resulting model.

To avoid such a difficulty, Theil has proposed a criterion that takes into account the number of estimated parameters. Such a criterion is called the *adjusted coefficient of determination* and is denoted $\bar{R}^2$. It is defined by

$$\bar{R}^2 = 1 - \frac{\|\hat{u}\|^2/(n-K)}{\|Y - \bar{Y}e\|^2/(n-1)}, \tag{22.46}$$

or equivalently, by

$$1 - \bar{R}^2 = (1 - R^2)\frac{n-1}{n-K}.$$

It is easy to see that $\bar{R}^2$ is always smaller or equal to R^2, and that the difference between these two measures increases as the number K of explanatory variables increases.

Note also that $\bar{R}^2$ can be written as

$$\bar{R}^2 = 1 - \frac{s^2}{\|Y - \bar{Y}e\|^2/(n-1)}, \tag{22.47}$$

where $s^2 = \|\hat{u}\|^2/(n-K)$ is the best quadratic unbiased estimator of σ^2. Hence a choice among competing linear models based on maximizing $\bar{R}^2$ is equivalent to a choice based on minimizing s^2.

c) Inconsistency of the $\bar{R}^2$ Criterion

It is, however, important to note that the model selection criterion $\bar{R}^2$ is inconsistent in the following sense. Let $\mathcal{M}_1$ be a (normal) linear model nested in another (normal) linear model $\mathcal{M}_2$. Then the probability of choosing $\mathcal{M}_1$ using $\bar{R}^2$, when $\mathcal{M}_1$ is correct, does not converge to 1 as n increases to $+\infty$. For let $\mathbf{X}_1$ and $\mathbf{X}_2$ be matrices of size $n \times K_1$ and $n \times K_2$ associated with the two models $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Because model $\mathcal{M}_1$ is nested in model $\mathcal{M}_2$, we have $\mathrm{Im}(\mathbf{X}_1) \subset \mathrm{Im}(\mathbf{X}_2)$ and $K_1 < K_2$, where $\mathrm{Im}(\mathbf{X})$ denotes the (image) space generated by the columns of $\mathbf{X}$. Then, if model $\mathcal{M}_1$ is correct, the probability of choosing $\mathcal{M}_1$ is

$$\begin{aligned} P(s_1^2 < s_2^2) &= P\left(\frac{\|\hat{u}_1\|^2}{n-K_1} < \frac{\|\hat{u}_2\|^2}{n-K_2}\right) \\ &= P\left(\frac{\|\hat{u}_1\|^2 - \|\hat{u}_2\|^2}{\sigma^2} < \frac{K_2 - K_1}{n-K_2}\frac{\|\hat{u}_2\|^2}{\sigma^2}\right). \end{aligned}$$

Under $\mathcal{M}_1$, the variable

$$U_n = \frac{\|\hat{u}_1\|^2 - \|\hat{u}_2\|^2}{\sigma^2}$$

has a constant probability distribution, namely a $\chi^2(K_2 - K_1)$ distribution, and hence converges in distribution to that distribution. On the other hand, the variable

$$V_n = \frac{K_2 - K_1}{n - K_2} \frac{\|\hat{u}_2\|^2}{\sigma^2}$$

converges in distribution to the constant $K_2 - K_1$ since the probability distribution of $\|\hat{u}_2\|^2/\sigma^2$ is a $\chi^2(n - K_2)$ distribution. Hence $P(s_1^2 < s_2^2)$ converges to $P(U < K_2 - K_1)$, where U is distributed as a $\chi^2(K_2 - K_1)$. Such a limit is strictly smaller than one.

22.3.2 The Akaike Criterion

a) The Maximum Log-likelihood

In Section 22.1.1 we saw that the asymptotic Kullback proximity between the true probability distribution and a given parametric model is

$$K = \min_{\theta} E_x E_0 \log \frac{f_0(Y \mid X)}{f(Y \mid X; \theta)} = E_x E_0 \log \frac{f_0(Y \mid X)}{f(Y \mid X; \theta^*)}, \quad (22.48)$$

where $f_0(Y \mid X)$ denotes the true conditional density of Y_i given X_i, and θ^* is the pseudo true value of the parameter θ. Thus a natural criterion for choosing among competing models can be based on the smallest value of K. This is, however, unfeasible since K is unknown. Hence, for a given model, we can think of replacing K by the quantity

$$\hat{K} = E_x E_0 \log \frac{f_0(Y \mid X)}{f(Y \mid X; \hat{\theta}_n)}, \quad (22.49)$$

where $\hat{\theta}_n$ is the maximum likelihood estimator of θ since $\hat{\theta}_n$ converges to θ^*. Note that $\hat{\theta}_n$ is considered as nonrandom in the expectations appearing in the expression (22.49).

The quantity $\hat{K}$ cannot be used since f_0 and the probability distribution of X are unknown. A natural solution consists of approximating $\hat{K}$ by $\tilde{K}$ defined as

$$\tilde{K} = \frac{1}{n} \sum_{i=1}^{n} \log f_0(y_i \mid x_i) - \frac{1}{n} \sum_{i=1}^{n} \log f(y_i \mid x_i; \hat{\theta}_n). \quad (22.50)$$

Though $\tilde{K}$ converges to K, as desired, $\tilde{K}$ cannot still be used because f_0 is unknown. Nonetheless, it is useful to note that the first term of $\tilde{K}$, namely $(1/n)\sum_{i=1}^{n} \log f_0(y_i \mid x_i)$ remains constant when comparing two competing models. Therefore minimizing $\tilde{K}$ is equivalent to maximizing over competing models the maximum of the log-likelihood

$$\sum_{i=1}^{n} \log f(y_i \mid x_i; \hat{\theta}_n).$$

Such a criterion is clearly not fully satisfying because it shares the same drawback as the coefficient of determination R^2, namely it increases as soon as one considers a more general model.

b) Definition of the Akaike Criterion

To solve such a difficulty, Akaike has proposed a method leading to a criterion whose practical performances seem to be adequate. The foundations of the Akaike criterion, however, are not fully satisfactory. Its derivation is as follows.

Suppose that the true probability density $f_0(y \mid x)$ belongs to the specified family $\{f(y \mid x; \theta), \theta \in \Theta\}$. Let θ_0 denote the value of θ associated to f_0. Under such an assumption, θ^* is equal to θ_0, K is equal to zero, and both $\hat{K}$ and $\tilde{K}$ converge to zero. It follows that the mean of the asymptotic distribution of the difference $\tilde{K} - \hat{K}$ is zero. Then, Akaike's idea consists in modifying $\tilde{K}$ so that the mean of the asymptotic distribution of $n(\tilde{K} - \hat{K})$ is also zero. That is, one seeks to eliminate the bias of order $1/n$.

We have

$$n\tilde{K} = \sum_{i=1}^{n} \log f(y_i \mid x_i; \theta_0) - \sum_{i=1}^{n} \log f(y_i \mid x_i; \hat{\theta}_n). \tag{22.51}$$

From a Taylor series expansion of the first term around $\hat{\theta}_n$, this quantity, which is always nonpositive, is asymptotically equivalent to

$$-\frac{n}{2}(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0), \tag{22.52}$$

where $\mathcal{I}(\theta_0)$ is the Fisher information matrix. Thus the asymptotic distribution of $-2n\tilde{K}$ is a $\chi^2(p)$ distribution, where p is the dimension of θ. Moreover, the asymptotic expectation of $n\tilde{K}$ is $-p/2$.

Consider now $n\hat{K}$. We have

$$n\hat{K} = n(E_x E_0 \log f(y \mid x; \theta_0) - E_x E_0 \log f(y \mid x; \hat{\theta}_n)).$$

From a Taylor expansion series of the second term around θ_0, this quantity, which is always nonnegative, is asymptotically equivalent to

$$\begin{aligned} & n\left(-E_x E_0 \frac{\partial \log f(Y \mid X; \theta_0)}{\partial \theta}(\hat{\theta}_n - \theta_0)\right. \\ & \qquad \left. -\frac{1}{2}(\hat{\theta}_n - \theta_0)' E_x E_0 \frac{\partial^2 \log f(Y \mid X; \theta_0)}{\partial \theta \partial \theta'}(\hat{\theta}_n - \theta_0)\right) \\ = \; & \frac{n}{2}(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \end{aligned} \tag{22.53}$$

since $E_x E_0(\partial \log f(Y \mid X; \theta_0)/\partial \theta) = 0$. Thus the asymptotic expectation of $n\hat{K}$ is $p/2$.

Replacing $\tilde{K}$ by $\tilde{K}^* = \tilde{K} + p/n$, the preceding results imply that the asymptotic expectation of $n(\tilde{K}^* - \hat{K})$ is zero, as desired. Therefore the Akaike criterion consists in selecting the model that provides the minimum value of

$$\tilde{K}^* = \frac{1}{n}\sum_{i=1}^{n} \log f_0(y_i \mid x_i) - \frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid x_i; \hat{\theta}_n) + \frac{p}{n},$$

or equivalently, the model that provides the maximum value of

$$AIC = \sum_{i=1}^{n} \log f(y_i \mid x_i; \hat{\theta}_n) - p = L_n(\hat{\theta}_n) - p, \tag{22.54}$$

where $L_n(\theta)$ denotes the (conditional) log-likelihood of the model. In other words, the so-called *Akaike information criterion* (AIC) is the maximum log-likelihood penalized by a quantity equal to the number of parameters.

c) Application to Linear Models

We consider again the normal linear model

$$Y = \mathbf{X}b + u, \quad u \sim N(0, \sigma^2 \mathbf{I}),$$

where $\mathbf{X}$ is a $n \times K$ matrix of rank K.

The log-likelihood function is

$$-\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(Y - \mathbf{X}b)'(Y - \mathbf{X}b).$$

Its maximum is

$$L_n(\hat{\theta}_n) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2},$$

where $\hat{\sigma}^2 = \|\hat{u}\|^2/n$.

Thus the Akaike information criterion is

$$AIC = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2} - (K+1).$$

Hence, when several competing normal linear models with different matrices $\mathbf{X}_j$ with ranks K_j, $j = 1, \ldots, J$, are considered, then one selects the model that gives the maximum value of

$$AIC^* = -\log\hat{\sigma}_j^2 - 2\frac{K_j+1}{n},$$

or equivalently, the model that gives the minimum value of

$$\log\hat{\sigma}_j^2 + 2(K_j+1)/n.$$

As for the coefficient of determination, the Akaike information criterion is inconsistent in the following sense. Suppose that one considers two nested normal linear models with K_1 and K_2 explanatory variables. Let $\mathcal{M}_1 \subset \mathcal{M}_2$ so that $(K_1 < K_2)$. Then the Akaike information criterion leads to preferring model $\mathcal{M}_1$ over model $\mathcal{M}_2$ if

$$\log\frac{\|\hat{u}_1\|^2}{n} + 2\frac{K_1+1}{n} < \log\frac{\|\hat{u}_2\|^2}{n} + 2\frac{K_2+1}{n},$$

i.e., if

$$\log\frac{\|\hat{u}_1\|^2}{\|\hat{u}_2\|^2} \leq 2\frac{(K_2-K_1)}{n},$$

i.e., if

$$\|\hat{u}_1\|^2 \leq \|\hat{u}_2\|^2 \exp\left(\frac{2(K_2-K_1)}{n}\right),$$

i.e., if

$$\frac{\|\hat{u}_1\|^2 - \|\hat{u}_2\|^2}{\sigma^2} \leq \frac{\|\hat{u}_2\|^2}{\sigma^2 n} n\left(\exp\frac{2(K_2-K_1)}{n} - 1\right).$$

Now, the term $n(\exp(2(K_2-K_1)/n - 1))$ converges to a constant, namely to $2(K_2-K_1)$ as n increases to $+\infty$. Hence the same argument as the one used for the adjusted coefficient of determination shows that the probability of choosing $\mathcal{M}_1$, when $\mathcal{M}_1$ is true, converges to a number that

is strictly smaller than 1. Therefore Akaike's procedure is inconsistent when $\mathcal{M}_1$ is true.

On the other hand, suppose that the true distribution belongs to $\mathcal{M}_2 - \mathcal{M}_1$. Then we know that $(\|\hat{u}_1\|^2 - \|\hat{u}_2\|^2)/n$ converges to a strictly positive constant. Hence the left-hand side of the inequality

$$\frac{\|\hat{u}_1\|^2 - \|\hat{u}_2\|^2}{\sigma^2} \leq \frac{\|\hat{u}_2\|^2}{\sigma^2 n}.n\left(\exp\frac{2(K_2 - K_1)}{n} - 1\right)$$

converges to $+\infty$, while the right-hand side converges to a constant. Therefore the probability of accepting $\mathcal{M}_2$, when $\mathcal{M}_2 - \mathcal{M}_1$ is true, converges to 1.

That is, inconsistency of Akaike's procedure only arises when the smaller hypothesis is true.

c) Schwarz Criterion

Various modifications of the Akaike criterion have been proposed. For instance, using Bayesian arguments, Schwarz has proposed the use of the following criterion

$$BIC = L_n(\hat{\theta}_n) - \frac{p}{2}\log n. \tag{22.55}$$

Thus the penalty associated with the number of estimated parameters is stronger in the Schwarz criterion than in the Akaike criterion as soon as $\log n$ is larger than 2, i.e., as soon as the number of observations n is larger than 8.

For normal linear models, the Schwarz criterion reduces to minimizing

$$BIC^* = \log\hat{\sigma}_j^2 + \frac{K_j + 1}{n}\log n. \tag{22.56}$$

Unlike previous criteria, the Schwarz criterion has the advantage of being consistent. For instance, consider the case of two competing nested normal linear models $\mathcal{M}_1 \subset \mathcal{M}_2$. Then $\mathcal{M}_1$ is preferred to $\mathcal{M}_2$ if

$$\frac{\|\hat{u}_1\|^2 - \|\hat{u}_2\|^2}{\sigma^2} \leq \frac{\|\hat{u}_2\|^2}{\sigma^2 n}.n\left(\exp\left(\frac{2(K_2 - K_1)}{n}\log n\right) - 1\right),$$

i.e., if

$$\frac{\|\hat{u}_1\|^2 - \|\hat{u}_2\|^2}{\sigma^2 \log n} \leq \frac{\|\hat{u}_2\|^2}{\sigma^2 n}\frac{n}{\log n}\left(\exp\left(\frac{K_2 - K_1}{n}\log n\right) - 1\right).$$

Under $\mathcal{M}_1$, the left-hand side converges in distribution to zero as n increases to $+\infty$, while the right-hand side converges in distribution to a constant, namely to $K_2 - K_1 > 0$. It follows that, under $\mathcal{M}_1$, the probability of selecting $\mathcal{M}_1$ converges to 1 and, hence, the Schwarz procedure is consistent. On the other hand, under $\mathcal{M}_2 - \mathcal{M}_1$, as for the Akaike criterion, it is easy to verify that the left-hand side diverges to $+\infty$, while the right-hand side converges to $K_2 - K_1$. Hence the probability of accepting $\mathcal{M}_2$ converges to one, as desired.

22.3.3 Bayesian Criteria

Let $\mathcal{M}_j$, $j = 1, \ldots, J$, be J competing models. Let $\ell_j(y \mid x; \theta)$ denote the likelihood function asociated with model $\mathcal{M}_j$, where $\theta \in \Theta_j$. It is assumed that a prior probability distribution on $\{1, \ldots, J\}$ is available with mass probabilities π_j, $j = 1, \ldots, J$. Conditionally upon $\mathcal{M}_j$, the parameter vector θ has a prior probability distribution defined on Θ_j with density $\pi(\theta \mid \mathcal{M}_j)$ with respect to some measure μ_j.

Let $\ell(y \mid x; \mathcal{M}_j)$ denote the conditional probability density of Y given X and $\mathcal{M}_j$. This density is obtained by integrating out the parameter θ, i.e.

$$\ell(y \mid x; \mathcal{M}_j) = \int_\Theta \ell_j(y \mid x; \theta)\pi(\theta \mid \mathcal{M}_j)d\mu_j(\theta). \tag{22.57}$$

Let $L(d, e)$ denote the loss incurred when one chooses the model $d \in \{1, \ldots, J\}$ when the true state of nature, i.e., the true model, is $e \in \{1, \ldots, J\}$. To a nonrandomized decision rule, i.e., a function defined on the sample space $\mathcal{Y}$ with values in $\{1, \ldots, J\}$, we can associate the Bayesian risk

$$R(\delta) = \sum_{j=1}^{J} \pi_j \int_{\mathcal{Y}} L(\delta(y), j)\ell(y \mid x; \mathcal{M}_j)d\nu(y). \tag{22.58}$$

Then we seek the decision rule δ that minimizes the risk $R(\delta)$. To do so, we note that $R(\delta)$ can be rewritten as

$$R(\delta) = \int_{\mathcal{Y}} \ell(y \mid x) \sum_{j=1}^{J} L(\delta(y), j)\pi(j \mid y)d\nu(y), \tag{22.59}$$

where

$$\ell(y \mid x) = \sum_{j=1}^{J} \ell(y \mid x; \mathcal{M}_j)\pi_j$$

and

$$\pi(j \mid y) = \frac{\ell(y \mid x; \mathcal{M}_j)\pi_j}{\ell(y \mid x)}.$$

To simplify the notation, we have omitted the conditioning on x in $\pi(j \mid y)$. Equation (22.59) shows that an optimal decision rule δ is given by

$$\delta(y) = \text{argmin}_d \sum_{j=1}^{J} L(d, j)\pi(j \mid y). \tag{22.60}$$

Hence we have established the following property.

Property 22.6: *Given an observation y, the optimal Bayesian decision rule is to choose the model d that minimizes the a posteriori average loss $\sum_{j=1}^{J} L(d, j)\pi(j \mid y)$.*

Frequently one chooses a loss function satisfying

$$\begin{aligned} L(i, i) &= 0, \\ L(i, j) &= a > 0, \quad i \neq j. \end{aligned}$$

Then the a posteriori average loss associated with a decision rule d is $a(1 - \pi(d \mid y))$. Hence, according to the Bayesian approach, one selects the model d that maximizes the posterior probability $\pi(d \mid y)$.

The main difficulty with the Bayesian approach, however, lies in the specification of the prior distributions on the parameters and the prior probabilities associated with the competing models. Such a specification is especially difficult in econometrics.

22.4 Appendix 1

ASYMPTOTIC DISTRIBUTIONS OF $\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))$, $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$, $(1/\sqrt{n})\hat{\lambda}_n^1$ AND $(1/\sqrt{n})\hat{\lambda}_n^2$ UNDER H_g.

(i) ASYMPTOTIC DISTRIBUTIONS OF $\sqrt{n}((\hat{\alpha}_n - \alpha_0)', (\hat{\beta}_n - b(\alpha_0))')'$ UNDER H_g:

The first-order Taylor series expansions of $(1/\sqrt{n})(\partial L_n^g(\hat{\alpha}_n)/\partial\alpha) = 0$ and $(1/\sqrt{n})(\partial L_n^h(\hat{\beta}_n)/\partial\beta) = 0$ around α_0 and $b(\alpha_0)$, respectively, give

$$\begin{aligned} \sqrt{n}(\hat{\alpha}_n - \alpha_0) &\# \mathbf{C}_{gg}^{-1} \frac{1}{\sqrt{n}} \frac{\partial L_n^g(\alpha_0)}{\partial \alpha_0}, \\ \sqrt{n}(\hat{\beta}_n - b(\alpha_0)) &\# \mathbf{K}_h^{-1} \frac{1}{\sqrt{n}} \frac{\partial L_n^h(b(\alpha_0))}{\partial \beta}, \end{aligned}$$

where the symbol # means, as before, that the difference between the left-hand side and right-hand side converges in probability to zero. It follows that

$$\sqrt{n}\begin{pmatrix} \hat{\alpha}_n - \alpha_0 \\ \hat{\beta}_n - b(\alpha_0) \end{pmatrix} \xrightarrow{d} N(0, \mathbf{\Omega\Sigma\Omega}),$$

where

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{C}_{gg}^{-1} & 0 \\ 0 & \mathbf{K}_h^{-1} \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{C}_{gg} & \mathbf{C}_{gh} \\ \mathbf{C}_{hg} & \mathbf{C}_{hh} \end{pmatrix}.$$

(ii) ASYMPTOTIC DISTRIBUTION OF $\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))$ UNDER H_g:
We have

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n)) &= \sqrt{n}(\hat{\beta}_n - b(\alpha_0)) - \sqrt{n}(b(\hat{\alpha}_n) - b(\alpha_0)) \\ &\# \sqrt{n}(\hat{\beta}_n - b(\alpha_0)) - \frac{\partial b(\alpha_0)}{\partial \alpha'}\sqrt{n}(\hat{\alpha}_n - \alpha_0) \\ &= \left(-\frac{\partial b(\alpha_0)}{\partial \alpha'}, \mathbf{I}\right)\sqrt{n}\begin{pmatrix} \hat{\alpha}_n - \alpha_0 \\ \hat{\beta}_n - b(\alpha_0) \end{pmatrix}. \end{aligned}$$

We now evaluate $\partial b(\alpha_0)/\partial \alpha'$. By definition, for any given α, $b(\alpha)$ is obtained by maximizing the quantity

$$E_x E_\alpha \log h(Y \mid X; \beta)$$

with respect to β. Hence we have

$$E_x E_\alpha \frac{\partial \log h(Y \mid X; b(\alpha))}{\partial \beta} = 0.$$

Then, differentiating this equality with respect α, we obtain

$$\begin{aligned} \frac{\partial b(\alpha_0)}{\partial \alpha'} &= -\left(E_x E_\alpha \frac{\partial^2 \log h(Y \mid X; b(\alpha))}{\partial \beta \partial \beta'}\right)^{-1} \\ &\quad \times E_x E_\alpha \left(\frac{\partial \log h(Y \mid X; b(\alpha))}{\partial \beta}\frac{\partial \log g(Y \mid X; \alpha)}{\partial \alpha'}\right). \end{aligned}$$

Hence

$$\frac{\partial b(\alpha_0)}{\partial \alpha'} = \mathbf{K}_h^{-1}\mathbf{C}_{hg}.$$

Using this expression in that of $\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))$, we obtain

$$\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n)) \# (-\mathbf{K}_h^{-1}\mathbf{C}_{hg}, \mathbf{I})\sqrt{n}\begin{pmatrix} \hat{\alpha}_n - \alpha_0 \\ \hat{\beta}_n - b(\alpha_0) \end{pmatrix}.$$

It follows that $\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))$ is asymptotically normally distributed with zero mean and variance covariance matrix

$$\begin{aligned} V_{as}(\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))) &= (-\mathbf{K}_h^{-1}\mathbf{C}_{hg}, \mathbf{I})\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Omega}\begin{pmatrix} -\mathbf{C}_{gh}\mathbf{K}_h^{-1} \\ \mathbf{I} \end{pmatrix} \\ &= \mathbf{K}_h^{-1}(\mathbf{C}_{hh} - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh})\mathbf{K}_h^{-1} \\ &= \mathbf{V}. \end{aligned}$$

(iii) ASYMPTOTIC DISTRIBUTION OF $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$ UNDER H_g:
We have

$$\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n)) = \sqrt{n}(\hat{\beta}_n - b(\alpha_0)) - \sqrt{n}(b_n(\hat{\alpha}_n) - b(\alpha_0)).$$

First, we shall find a quantity that is asymptotically equivalent to $\sqrt{n}(b_n(\hat{\alpha}_n) - b(\alpha_0))$. Consider the function $\phi_n(\alpha, \beta)$ defined by

$$\phi_n(\alpha, \beta) = \sum_{i=1}^{n} E_\alpha \frac{\partial \log h(Y \mid X_i; \beta)}{\partial \beta}.$$

By definition of $b_n(\alpha)$, we have

$$\phi_n(\alpha, b_n(\alpha)) = 0, \quad \forall \alpha.$$

In particular

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\phi_n(\hat{\alpha}_n, b_n(\hat{\alpha}_n)) \\ &\# \frac{1}{\sqrt{n}}\phi_n(\alpha_0, b(\alpha_0)) + \frac{1}{n}\frac{\partial \phi_n(\alpha_0, b(\alpha_0))}{\partial \alpha'}\sqrt{n}(\hat{\alpha}_n - \alpha_0) \\ &\quad + \frac{1}{n}\frac{\partial \phi_n(\alpha_0, b(\alpha_0))}{\partial \beta'}\sqrt{n}(b_n(\hat{\alpha}_n) - b(\alpha_0)), \end{aligned}$$

which implies that

$$\begin{aligned} \sqrt{n}(b_n(\hat{\alpha}_n) - b(\alpha_0)) \# &\left(-\frac{1}{n}\frac{\partial \phi_n(\alpha_0, b(\alpha_0))}{\partial \beta'}\right)^{-1} \\ &\times \left(\frac{1}{\sqrt{n}}\phi_n(\alpha_0, b(\alpha_0)) + \frac{1}{n}\frac{\partial \phi_n(\alpha_0, b(\alpha_0))}{\partial \alpha'}\sqrt{n}(\hat{\alpha}_n - \alpha_0)\right), \end{aligned}$$

i.e.

$$\begin{aligned} &\sqrt{n}(b_n(\hat{\alpha}_n) - b(\alpha_0)) \\ &\quad \# \mathbf{K}_h^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E_{\alpha_0}\frac{\partial \log h(Y \mid X_i; b(\alpha_0))}{\partial \beta} + \mathbf{C}_{hg}\sqrt{n}(\hat{\alpha}_n - \alpha_0)\right). \end{aligned}$$

Next, using the latter asymptotic equivalence in the expression for $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$, we obtain

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n)) \quad \# \quad & \sqrt{n}(\hat{\beta}_n - b(\alpha_0)) \\ & -\mathbf{K}_h^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E_{\alpha_0}\frac{\partial \log h(Y \mid X_i; b(\alpha_0))}{\partial \beta} \\ & -\mathbf{K}_h^{-1}\mathbf{C}_{hg}\sqrt{n}(\hat{\alpha}_n - \alpha_0).\end{aligned}$$

It remains to determine the asymptotic distribution of the vector

$$\sqrt{n}\begin{pmatrix} \hat{\alpha}_n - \alpha_0 \\ \hat{\beta}_n - b(\alpha_0) \\ \frac{1}{n}\sum_{i=1}^{n} E_{\alpha_0}\frac{\partial \log h(Y \mid X_i; b(\alpha_0))}{\partial \beta} \end{pmatrix}.$$

This vector is asymptotically equivalent to

$$\gamma_n = \begin{pmatrix} \mathbf{C}_{gg}^{-1} & 0 & 0 \\ 0 & \mathbf{K}_h^{-1} & 0 \\ 0 & 0 & \mathbf{I} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log g(Y_i \mid X_i; \alpha_0)}{\partial \alpha} \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log h(Y_i \mid X_i; b(\alpha_0))}{\partial \beta} \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E_{\alpha_0}\frac{\partial \log h(Y_i \mid X_i; b(\alpha_0))}{\partial \beta} \end{pmatrix}.$$

Hence $\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n))$ is asymptotically equivalent to

$$\begin{aligned}&(-\mathbf{K}_h^{-1}\mathbf{C}_{hg}, \mathbf{I}, -\mathbf{K}_h^{-1})\gamma_n \\ &= \mathbf{K}_h^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial \log h(Y_i \mid X_i; b(\alpha_0))}{\partial \beta} - E_{\alpha_0}\frac{\partial \log h(Y_i \mid X_i; b(\alpha_0))}{\partial \beta}\right)\right. \\ &\quad \left. - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log g(Y_i \mid X_i; \alpha_0)}{\partial \alpha}\right).\end{aligned}$$

This vector is asymptotically normally distributed with zero mean and variance covariance matrix

$$\mathbf{V}^* = \mathbf{K}_h^{-1}(\mathbf{C}_{hh}^* - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh})\mathbf{K}_h^{-1}.$$

(iv) ASYMPTOTIC DISTRIBUTIONS OF $(1/\sqrt{n})\hat{\lambda}_n^1$ AND $(1/\sqrt{n})\hat{\lambda}_n^2$ UNDER H_g:

We have

$$\begin{aligned}\frac{1}{\sqrt{n}}\hat{\lambda}_n^1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log h(Y_i \mid X_i; b(\hat{\alpha}_n))}{\partial \beta} \\ &\# \frac{1}{\sqrt{n}}\frac{\partial L_n^h(b(\alpha_0))}{\partial \beta} - \mathbf{K}_h\sqrt{n}(b(\hat{\alpha}_n) - b(\alpha_0)) \\ &\# \mathbf{K}_h\sqrt{n}(\hat{\beta}_n - b(\alpha_0)) - \mathbf{K}_h\sqrt{n}(b(\hat{\alpha}_n) - b(\alpha_0)) \\ &= \mathbf{K}_h\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n)).\end{aligned}$$

Hence $(1/\sqrt{n})\hat{\lambda}_n^1$ is asymptotically normally distributed with zero mean and variance covariance matrix

$$\mathbf{W} = \mathbf{C}_{hh} - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh}.$$

Similarly, it can be shown that

$$\frac{1}{\sqrt{n}}\hat{\lambda}_n^2 \# \mathbf{K}_h\sqrt{n}(\hat{\beta}_n - b_n(\hat{\alpha}_n)).$$

Hence $(1/\sqrt{n})\hat{\lambda}_n^2$ is asymptotically normally distributed with zero mean and variance covariance matrix

$$\mathbf{W}^* = \mathbf{C}_{hh}^* - \mathbf{C}_{hg}\mathbf{C}_{gg}^{-1}\mathbf{C}_{gh}.$$

22.5 Appendix 2

ASYMPTOTIC DISTRIBUTION OF THE COX TEST STATISTIC.

We have

$$\sqrt{n}C_n = \frac{1}{\sqrt{n}}(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{\beta}_n)) - \sqrt{n}H(\hat{\alpha}_n).$$

(i) AN EXPANSION OF $\sqrt{n}H(\hat{\alpha}_n)$:

Let α_0 denote the true value of α under H_g. Then, under H_g, we have

$$\sqrt{n}H(\hat{\alpha}_n) \# \sqrt{n}H(\alpha_0) + \frac{\partial H(\alpha_0)}{\partial \alpha'}\sqrt{n}(\hat{\alpha}_n - \alpha_0),$$

where the symbol # means, as before, that the difference between the left-hand side and the right-hand side converges in probability to zero.

Now, the vector of derivatives $\partial H(\alpha_0)/\partial\alpha$ can be written as

$$\begin{aligned}\frac{\partial H(\alpha_0)}{\partial\alpha} &= E_x E_{\alpha_0}\left(\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right. \\ &\quad + \log g(Y \mid X;\alpha_0)\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha} \\ &\quad - \frac{\partial b'(\alpha_0)}{\partial\alpha}\frac{\partial \log h(Y \mid X;b(\alpha_0))}{\partial\beta} \\ &\quad \left. - \log h(Y \mid X;b(\alpha_0))\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right).\end{aligned}$$

Since α_0 and $b(\alpha_0)$ maximize $E_x E_{\alpha_0} \log g(Y \mid X;\alpha)$ and $E_x E_{\alpha_0} \log h(Y \mid X;\beta)$, respectively, we have

$$E_x E_{\alpha_0}\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha} = 0 \quad \text{and} \quad E_x E_{\alpha_0}\frac{\partial \log h(Y \mid X;b(\alpha_0))}{\partial\beta} = 0.$$

Hence

$$\frac{\partial H(\alpha_0)}{\partial\alpha} = E_x E_{\alpha_0}\left(\left(\log g(Y \mid X;\alpha_0) - \log h(Y \mid X;b(\alpha_0))\right)\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right).$$

Thus

$$\frac{\partial H(\alpha_0)}{\partial\alpha} = \operatorname{cov}\left(\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}, \log\frac{g(Y \mid X;\alpha_0)}{h(Y \mid X;b(\alpha_0))}\right) \tag{22.61}$$

since

$$E_x E_{\alpha_0}\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha} = 0.$$

On the other hand, from maximum likelihood theory (see Chapter 17), we have

$$\begin{aligned}\sqrt{n}(\hat{\alpha}_n - \alpha_0) \quad \# \quad & \left(E_x E_{\alpha_0}\left(\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha'}\right)\right)^{-1} \\ & \times \frac{1}{\sqrt{n}}\frac{\partial L_n^g(\alpha_0)}{\partial\alpha}.\end{aligned}$$

Hence we obtain

$$\begin{aligned}\sqrt{n}H(\hat{\alpha}_n) \;\#\; & \sqrt{n}H(\alpha_0) \\ & + \operatorname{cov}\left(\log\frac{g(Y \mid X;\alpha_0)}{h(Y \mid X;b(\alpha_0))}, \frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right) \\ & \times \left(V\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right)^{-1}\frac{1}{\sqrt{n}}\frac{\partial L_n^g(\alpha_0)}{\partial\alpha}.\end{aligned}$$

(ii) AN EXPANSION OF $(1/\sqrt{n})(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{\beta}_n))$:
We have

$$\begin{aligned}\frac{1}{\sqrt{n}}(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{\beta}_n)) \;=\; & \frac{1}{\sqrt{n}}(L_n^g(\alpha_0) - L_n^h(b(\alpha_0)) \\ & + \frac{1}{\sqrt{n}}(L_n^g(\hat{\alpha}_n) - L_n^g(\alpha_0)) \\ & + \frac{1}{\sqrt{n}}(L_n^h(\hat{\beta}_n) - L_n^h(b(\alpha_0)).\end{aligned}$$

But the Taylor series expansion of $L_n^g(\hat{\alpha}_n) - L_n^g(\alpha_0)$ (see Chapter 17) shows that

$$L_n^g(\hat{\alpha}_n) - L_n^g(\alpha_0) \;\#\; -\frac{1}{2}(\hat{\alpha}_n - \alpha_0)'\frac{\partial^2 L_n^g(\alpha_0)}{\partial\alpha\partial\alpha'}(\hat{\alpha}_n - \alpha_0).$$

Hence

$$\begin{aligned}&\frac{1}{\sqrt{n}}(L_n^g(\hat{\alpha}_n) - L_n^g(\alpha_0)) \\ &\quad\# \quad \frac{1}{2\sqrt{n}}\sqrt{n}(\hat{\alpha}_n - \alpha_0)'\left(-\frac{1}{n}\frac{\partial^2 L_n^g(\alpha_0)}{\partial\alpha\partial\alpha'}\right)\sqrt{n}(\hat{\alpha}_n - \alpha_0) \\ &\quad\# \quad 0.\end{aligned}$$

Similarly, it can be shown that $(1/\sqrt{n})(L_n^h(\hat{\beta}_n) - L_n^h(b(\alpha_0)))$ converges in probability to zero. Therefore, collecting results, we obtain

$$\frac{1}{\sqrt{n}}(L_n^g(\hat{\alpha}_n) - L_n^h(\hat{\beta}_n)) \;\#\; \frac{1}{\sqrt{n}}\Big(L_n^g(\alpha_0) - L_n^h(b(\alpha_0))\Big).$$

(iii) AN EXPANSION OF $\sqrt{n}C_n$:
We have

$$\begin{aligned}\sqrt{n}C_n \quad \# \quad & \frac{1}{\sqrt{n}}\left(L_n^g(\alpha_0) - L_n^h(b(\alpha_0))\right) - \sqrt{n}H(\alpha_0) \\ & -\text{cov}\left(\log\frac{g(Y \mid X;\alpha_0)}{h(Y \mid X;b(\alpha_0))}, \frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right) \\ & \times\left(V\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right)^{-1}\frac{1}{\sqrt{n}}\frac{\partial L_n^g(\alpha_0)}{\partial\alpha}.\end{aligned}$$

Hence

$$\begin{aligned}\sqrt{n}C_n \quad \# \quad & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\log\frac{g(Y_i \mid X_i;\alpha_0)}{h(Y_i \mid X_i;b(\alpha_0))} - E_x E_{\alpha_0}\log\frac{g(Y \mid X;\alpha_0)}{h(Y \mid X;b(\alpha_0))}\right. \\ & -\text{cov}\left(\log\frac{g(Y \mid X;\alpha_0)}{h(Y \mid X;b(\alpha_0))}, \frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right) \\ & \left.\times\left(V\frac{\partial \log g(Y \mid X;\alpha_0)}{\partial\alpha}\right)^{-1}\frac{\partial \log g(Y_i \mid X_i;\alpha_0)}{\partial\alpha}\right).\end{aligned}$$

(iv) ASYMPTOTIC DISTRIBUTION OF $\sqrt{n}C_n$:
The asymptotic equivalence obtained in (iii) implies that the asymptotic distribution of $\sqrt{n}C_n$ under H_g is that of $(1/\sqrt{n})\sum_{i=1}^{n} Z_i$, where Z_i is the i-th residual in the linear regression with constant term of $\log(g(Y_i \mid X_i;\alpha_0)/h(Y_i \mid X_i;b(\alpha_0))$ on the components of $\partial \log g(Y_i \mid X_i;\alpha_0)/\partial\alpha$.

We can now apply a Central Limit Theorem since the Z_i's are independently and identically distributed with zero means. Moreover, the interpretation of Z_i as a residual in a linear regression shows that the variance of Z_i is equal to ω_0^2, as given in Property 22.3.

22.6 Exercises

EXERCISE 22.1: Consider the stationary autoregressive process of order 1 discussed in Section 22.1.2-b). Let

$$C_{2\to 1} = \frac{1}{2}\log\frac{V_0(Y_{1t} \mid Y_{1t-1})}{V_0(Y_{1t} \mid Y_{1t-1}, Y_{2t-1})}$$

be the measure of causality of Y_2 on Y_1.

a) By analogy to the measure $C_{2\to 1}$, show that one can define a measure of causality of Y_1 on Y_2, an instantaneous measure of causality between Y_1 and Y_2, and a measure of dependence between Y_1 and Y_2. Let these measures be denoted $C_{1\to 2}$, $C_{1\leftrightarrow 2}$, and $C_{1,2}$, respectively. Show that $C_{1,2} = C_{1\to 2} + C_{2\to 1} + C_{1\leftrightarrow 2}$.

b) Show that the preceding results can be generalized to a bivariate stationary autoregressive process of order p.

EXERCISE 22.2: Consider the framework of Section 22.2. Using the asymptotic efficiency of $b(\hat{\alpha}_n)$ under H_g, show that we have

$$V_{as}(\sqrt{n}(\hat{\beta}_n - b(\hat{\alpha}_n))) = V_{as}(\sqrt{n}(\hat{\beta}_n - b(\alpha_0))) - V_{as}(\sqrt{n}(b(\hat{\alpha}_n) - b(\alpha_0)))$$

under H_g.

EXERCISE 22.3: Using the results of Section 22.2.5, show how one can construct a test of the trichotomous logit model defined by

$$\begin{aligned}
\tilde{g}_1(x_i, \alpha) &= \frac{1}{1 + \exp x_{1i}\alpha_1 + \exp x_{2i}\alpha_2}, \\
\tilde{g}_2(x_i, \alpha) &= \frac{\exp x_{1i}\alpha_1}{1 + \exp x_{1i}\alpha_1 + \exp x_{2i}\alpha_2}, \\
\tilde{g}_3(x_i, \alpha) &= \frac{\exp x_{2i}\alpha_2}{1 + \exp x_{1i}\alpha_1 + \exp x_{2i}\alpha_2}
\end{aligned}$$

against the nested logit model defined by

$$\begin{aligned}
\tilde{h}_1(x_i, \beta) &= \frac{1}{1 + \exp x_{3i}\beta_1}, \\
\tilde{h}_2(x_i, \beta) &= \frac{\exp x_{3i}\beta_1}{1 + \exp x_{3i}\beta_1} \frac{1}{1 + \exp x_{4i}\beta_2}, \\
\tilde{h}_3(x_i, \beta) &= \frac{\exp x_{3i}\beta_1}{1 + \exp x_{3i}\beta_1} \frac{\exp x_{4i}\beta_2}{1 + \exp x_{4i}\beta_2}.
\end{aligned}$$

EXERCISE 22.4: Determine the form of the generalized score test based on $\xi_n^{S_2}$ when the two competing models H_g and H_h are exponential and given by

$$H_g : g(y_i \mid x_i; \alpha) = \exp\left(\sum_{h=1}^{G} A_h(\alpha) S_h(y_i, x_i) + B(x_i, \alpha) + C(y_i, x_i)\right)$$

and

$$H_h : h(y_i \mid x_i;\beta) = \exp\left(\sum_{h=1}^{H} \tilde{A}_h(\beta)\tilde{S}_h(y_i,x_i) + \tilde{B}(x_i,\beta) + \tilde{C}(y_i,x_i)\right).$$

EXERCISE 22.5: Consider the test statistics $\hat{\lambda}_n$ and T_λ given by equations (22.36) and (22.37), respectively. Assuming that the subspaces generated by the columns of $\mathbf{X}$ and $\mathbf{Z}$ are orthogonal, show that $\hat{\lambda}_n = 1$. Conclude that $T_\lambda^2 = Y'\mathbf{M}_z Y/\hat{\sigma}_n^2$ and that T_λ^2 is distributed as a $\chi^2(d)$ under H_g, where d is the dimension of z_i.

EXERCISE 22.6: Let $\bar{R}_K^2$ be the adjusted coefficient of determination in a linear model with K explanatory variables. Let $\bar{R}_{K-1}^2$ be the adjusted coefficient of determination in the model obtained by deleting one explanatory variable. Show that

$$\bar{R}_K^2 - \bar{R}_{K-1}^2 = \frac{1-\bar{R}_K^2}{n-K+1}(t^2-1),$$

where t is the Student statistic associated with the deleted explanatory variable.

22.7 References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory Society*, Petrov-Csaki ed., Akademiaio Kiado, Budapest.

Amemiya, T. (1980). "Selection of Regressors," *International Economic Review*, 21, 331-354.

Cox, D.R. (1961). "Tests of Separate Families of Hypotheses," in *Proceedings of the Fourth Berkeley Symposium*, Vol.1.

Cox, D.R. (1962). "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society*, 24, B, 406-424.

Davidson, R. and McKinnon, J.G. (1981). "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, 49, 781-793.

Gourieroux, C., Monfort, A., and Trognon, A. (1983). "Testing Nested or Non-Nested Hypotheses," *Journal of Econometrics*, 21, 83-115.

McKinnon, J.G., White, H., and Davidson, R. (1983). "Tests for Model Specifications in the Presence of Alternative Hypotheses: Some Further Results," *Journal of Econometrics*, 21, 53-70.

Pesaran, M.H. (1974). "On the General Problem of Model Selection," *Review of Economic Studies*, 41, 153-171.

Pesaran, M.H. and Deaton, A.S. (1978). "Testing Non Nested Non Linear Regression Models," *Econometrica*, 46, 677-694.

Sawa, T. (1978). "Information Criteria for Discrimination Among Alternative Regression Models," *Econometrica*, 46, 1273-1291.

Schwarz , G. (1978). "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.

CHAPTER 23

Asymptotic Efficiency

In preceding chapters, especially in Chapters 6, 15, and 16, we studied the *optimal* properties of general estimation and testing procedures in finite samples. On the other hand, in large samples the optimality problem was discussed in some special cases only. The goal of this chapter is to study such a problem more systematically.

The theory of asymptotic efficiency is technically difficult. It is also based on tools that are relatively different from those used up to now. For these two reasons we shall present the main results of this theory on an intuitive basis. In the same spirit of simplification the theoretical aspects will be discussed within a sampling framework. Then we shall consider successively the asymptotic optimality of estimators and the asymptotic optimality of tests. For each of these problems there exist different optimality concepts depending on the order of approximation retained in the asymptotic expansions. These concepts correspond to the properties of consistency, first-order asymptotic efficiency and second-order asymptotic efficiency.

23.1 Asymptotic Comparison of Estimators

Hereafter we consider a random sample $(Y_1, \ldots, Y_n)$ drawn from a common probability distribution that belongs to a family of probability distributions parameterized by a parameter vector $\theta \in \Theta$. We assume that the statistical model is homogenous and we let $f(y; \theta)$ denote the marginal probability density for one observation. The object of interest is the parameter vector θ.

It is natural to restrict a priori the study to estimators $T_n(Y)$ that are consistent for θ. There exist, however, an infinite number of consistent estimators. To choose among those estimators, it is therefore necessary

to invoke additional criteria based on the concept of asymptotic precision.

23.1.1 First-Order Asymptotic Efficiency

a) Three Alternative Definitions

There exist many definitions of asymptotic efficiency that can generalize the concept of efficiency introduced in finite samples. First, we note that the property of consistency is relatively close to that of asymptotic unbiasedness. We are thus led to compare (asymptotically) unbiased estimators.

(i) The concept of asymptotic efficiency, which was used up to now, is based on the asymptotic variance covariance matrix of the estimator. Let

$$\mathcal{I}_1(\theta) = E_\theta \left(-\frac{\partial^2 \log f(Y;\theta)}{\partial\theta\partial\theta'} \right)$$

denote the Fisher information matrix corresponding to one observation. We suppose that this matrix is nonsingular.

Definition 23.1: *A consistent estimator T_n of θ is first-order asymptotically efficient if*

$$V_{as}(\sqrt{n}(T_n(Y) - \theta)) = \mathcal{I}_1(\theta)^{-1}, \quad \forall\theta.$$

(ii) On the other hand, in finite samples we have seen that any efficient estimator, i.e., any unbiased estimator whose variance covariance matrix reaches the Frechet–Darmois–Cramer–Rao (FDCR) bound, is optimal in the class of unbiased estimators (see Chapter 6). Moreover, we know that there exist asymptotically efficient estimators in the sense of Definition 23.1. For instance, the maximum likelihood estimator is such an estimator. Thus any (asymptotically) optimal estimators in the class of (asymptotically) unbiased estimators should be in principle asymptotically efficient. This leads to the following alternative definition of first-order asymptotic efficiency.

Definition 23.2: *A consistent estimator T_n of θ is first-order asymptotically efficient if its asymptotic variance covariance matrix is minimal in the class of consistent estimators of θ.*

(iii) Lastly, we can introduce another definition based on the concept of information contained in the initial model and on the information contained in the model induced by the estimator T_n.

Specifically, let $\mathcal{I}^{T_n}(\theta)$ denote the Fisher information matrix associated with the model induced by T_n. If $\ell_n(T_n;\theta)$ denote the probability density of T_n, then we have

$$\mathcal{I}^{T_n}(\theta) = E_\theta\left(-\frac{\partial^2 \log \ell_n(T_n;\theta)}{\partial\theta\partial\theta'}\right) \preceq n\mathcal{I}_1(\theta)$$

using the properties of Fisher information. The next definition requires that the estimator T_n contains asymptotically the same information as the original data, i.e., that T_n is "asymptotically sufficient" in some sense.

Definition 23.3: *A consistent estimator T_n of θ is first-order asymptotically efficient if the information contained in T_n is asymptotically equal to the information contained in $Y_1, \ldots, Y_n$, i.e.*

$$\lim_{n\to\infty} \frac{1}{n}\mathcal{I}^{T_n}(\theta) = \mathcal{I}_1(\theta).$$

The preceding three definitions are given without the regularity conditions under which these definitions are meaningful (existence of the asymptotic variance covariance matrix, rate of convergence in $1/\sqrt{n}$ of the estimator, etc.). Such regularity conditions differ according to which definition is retained. As a result, these three natural definitions of first-order asymptotic efficiency are not, in general equivalent. It can even be proved that any one of them does not imply the other two without additional assumptions.

b) Equivalence of the Three Definitions

The equivalence between the three definitions given above hold provided one restricts the class of estimators that are considered. The additional assumption that we introduce is

Assumption 23.1: *the estimator T_n satisfies*

$$\begin{pmatrix} \sqrt{n}(T_n(Y)-\theta) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial \log f(Y_i;\theta)}{\partial\theta} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{A}_{11}(\theta) & \mathbf{A}_{12}(\theta) \\ \mathbf{A}_{21}(\theta) & \mathbf{A}_{22}(\theta) \end{pmatrix}\right)$$

as $n \to +\infty$, and this convergence is uniform in the parameter θ.

Assumption 23.1 is as analogous to the regularity condition that was introduced when deriving the FDCR bound. See Section 6.2.

Under this additional assumption and other more classical ones (in particular, the existence of the expectations of estimators), we can apply an argument similar to the one used in finite samples. Specifically, we have

$$\lim_{n\to\infty} E_\theta(T_n(Y) - \theta) = 0,$$

which is equivalent to

$$\lim_n \int_{\mathcal{Y}} T_n(y) \prod_{i=1}^n f(y_i;\theta) \bigotimes_{i=1}^n d\mu(y_i) = \theta.$$

Differentiating with respect to θ, we obtain

$$\lim_n \int_{\mathcal{Y}} T_n(y) \frac{\partial \left(\prod_{i=1}^n f(y_i;\theta)\right)}{\partial\theta} \bigotimes_{i=1}^n d\mu(y_i) = \mathbf{I},$$

which is equivalent to

$$\lim_n \int_{\mathcal{Y}} \sqrt{n}(T_n(y) - \theta) \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(y_i;\theta)}{\partial\theta} \prod_{i=1}^n f(y_i;\theta) \bigotimes_{i=1}^n d\mu(y_i) = \mathbf{I}$$

using the fact that the score vector has zero mean. The latter statement is clearly equivalent to

$$\lim_n \mathrm{Cov}_\theta \left(\sqrt{n}(T_n(Y) - \theta), \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(Y_i;\theta)}{\partial\theta} \right) = \mathbf{I}.$$

This establishes the next result.

Property 23.1: *Suppose that the estimator T_n satisfies Assumption 23.1. Then $\mathbf{A}_{12}(\theta) = \mathbf{I}$.*

Consider now the residual vector in the asymptotic or population linear regression of $\sqrt{n}(T_n(Y) - \theta)$ on $(1/\sqrt{n}) \sum_{i=1}^n (\partial \log f(Y_i;\theta)/\partial\theta)$, i.e.

$$S_n = \sqrt{n}(T_n(Y) - \theta) - \mathbf{A}_{22}(\theta)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(Y_i;\theta)}{\partial\theta}.$$

We have $V_{as} S_n = \mathbf{A}_{11}(\theta) - \mathbf{A}_{22}(\theta)^{-1}$. Noting that $\mathbf{A}_{22}(\theta) = \mathcal{I}_1(\theta)$, we obtain an inequality, which is the asymptotic analog of the FDCR inequality. This is formally stated in the next property.

Property 23.2: *Suppose that the estimator T_n satisfies Assumption 23.1. Then we have*

$$\mathbf{A}_{11}(\theta) = V_{as}(\sqrt{n}(T_n(Y) - \theta)) \succeq \mathcal{I}_1(\theta)^{-1}.$$

We can now established the desired equivalence.

Property 23.3: *Under the additional Assumption 23.1, Definitions 23.1 and 23.2 are equivalent.*

PROOF:

(i) The fact that Definition 23.1 implies Definition 23.2 is an immediate consequence of Property 23.2.

(ii) Conversely, because the maximum likelihood estimator satisfies Assumption 23.1 with $\mathbf{A}_{11}(\theta) = \mathcal{I}_1(\theta)^{-1}$, an estimator that is first-order asymptotically efficient in the sense of Definition 23.2 must have an asymptotic variance covariance matrix that is smaller or equal to $\mathcal{I}_1(\theta)^{-1}$. Therefore, using Property 23.2, it follows that this asymptotic variance covariance matrix must be equal to $\mathcal{I}_1(\theta)^{-1}$. □

The second desired equivalence holds, as the next property states.

Property 23.4: *Under the additional Assumption 23.1, Definition 23.3 is equivalent to the other two definitions.*

PROOF:

(i) We know that

$$n\mathcal{I}_1(\theta) = V_\theta\left(\sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta}\right)$$

and

$$\mathcal{I}^{T_n}(\theta) = V_\theta E_\theta\left(\sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta} \mid T_n(Y)\right).$$

See Appendix to Chapter 11. Thus the condition $\lim_{n\to\infty}(1/n)\mathcal{I}^{T_n}(\theta) = \mathcal{I}_1(\theta)$ is equivalent to

$$\lim_n \frac{1}{n} E_\theta V_\theta\left(\sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta} \mid T_n(Y)\right) = 0,$$

which is equivalent to

$$\lim_n E_\theta V_\theta \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta} \mid T_n(Y) \right) = 0,$$

i.e.

$$\lim_n E_\theta V_\theta \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta} \mid \sqrt{n}(T_n(Y) - \theta) \right) = 0.$$

Now, asymptotically $(1/\sqrt{n}) \sum_{i=1}^{n} (\partial \log f(Y_i;\theta)/\partial\theta)$ must be equal to its conditional expectation given $\sqrt{n}(T_n(Y) - \theta)$. Because asymptotic normality holds, the latter expectation must be equal to the population linear regression. Hence

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta} \# \mathbf{A}_{11}(\theta)^{-1} \sqrt{n}(T_n(Y) - \theta).$$

Then, taking the variance of each side, it follows that

$$\mathbf{A}_{22}(\theta) = \mathbf{A}_{11}(\theta)^{-1},$$

i.e.

$$\mathbf{A}_{11}(\theta) = (\mathbf{A}_{22}(\theta))^{-1} = \left(V_\theta \frac{\partial \log f(Y;\theta)}{\partial \theta} \right)^{-1} = \mathcal{I}_1(\theta)^{-1}.$$

Therefore Definition 23.3 implies Definition 23.1.

(ii) Conversely, if Definition 23.1 holds, then we have

$$\mathbf{A}_{22}(\theta) = \mathbf{A}_{11}(\theta)^{-1}.$$

Thus the residual vector

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta} - \mathbf{A}_{11}(\theta)^{-1} \sqrt{n}(T_n(Y) - \theta)$$

has a zero asymptotic variance covariance matrix. Hence the two terms are asymptotically linearly related. This implies

$$E_\theta \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta} \mid T_n(Y) \right) \# \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\theta)}{\partial \theta}.$$

Therefore

$$\frac{1}{n} \mathcal{I}^{T_n}(\theta) \# \mathcal{I}_1(\theta).$$

□

c) A Fourth Definition

The preceding discussion of first-order asymptotic efficiency suggests a fourth definition.

Definition 23.4: *Under Assumption 23.1, a consistent estimator T_n of θ is first-order asymptotically efficient if*

$$\sqrt{n}(T_n(Y) - \theta) \# a(\theta) + \mathbf{B}(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(Y_i;\theta)}{\partial \theta}.$$

From above we noted that the coefficients must necessarily satisfy $a(\theta) = 0$ and $\mathbf{B}(\theta) = \mathbf{A}_{22}(\theta)^{-1} = \mathcal{I}_1(\theta)^{-1}$.

Example 23.1: By definition, an M-estimator $\tilde{\theta}_n$ of the parameter vector θ is obtained by maximizing an objective function of the form $\sum_{i=1}^{n}\psi(Y_i;\theta)$. Hence it satisfies asymptotically

$$\sqrt{n}(\tilde{\theta}_n - \theta) \# \left(E_\theta\left(-\frac{\partial^2\psi(Y;\theta)}{\partial\theta\partial\theta'}\right)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\psi(Y_i;\theta)}{\partial\theta}.$$

Suppose now that such an estimation method is applied to a parametric model defined by the family $\{\ell(y;\theta) = \prod_{i=1}^{n} f(y_i;\theta), \theta \in \Theta\}$. Then Assumption 23.1 is satisfied and the M-estimator $\tilde{\theta}_n$ is first-order asymptotically efficient if there exists a matrix $\mathbf{B}(\theta)$ such that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\psi(Y_i;\theta)}{\partial\theta} \# \mathbf{B}(\theta)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\log f(Y_i;\theta)}{\partial\theta},$$

i.e.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial\psi(Y_i;\theta)}{\partial\theta} - \mathbf{B}(\theta)\frac{\partial\log f(Y_i;\theta)}{\partial\theta}\right) \# 0.$$

Using the fact that the Y_i's are independent and identically distributed, it is easy to see that this implies

$$\frac{\partial\psi(Y;\theta)}{\partial\theta} = \mathbf{B}(\theta)\frac{\partial\log f(Y;\theta)}{\partial\theta}.$$

This condition for first-order asymptotic efficiency can be related to the condition for consistency of M-estimators. Recall that M-estimators

are typically used when the parameter vector θ is defined by some estimating equations. As it is well-known, however, that there exist numerous choices for function g satisfying $E_{\theta_0} g(Y;\theta) = 0 \Rightarrow \theta = \theta_0$. To each of these choices, there corresponds a different interpretation of the parameter vector θ, in general. Moreover, as a consequence of Lemma 8.1, we know that an M-estimator is consistent if

$$\frac{\partial \psi(Y;\theta)}{\partial \theta} = \Lambda(\theta)\, g(Y;\theta),$$

where g is one of the functions defining the estimating equations.

In the present case it follows from the zero expectation of the score vector, i.e.

$$E_{\theta_0} \frac{\partial \log f(Y;\theta_0)}{\partial \theta} = 0,$$

that it is possible to choose the estimating equations based on

$$g(y;\theta) = \frac{\partial \log f(y;\theta)}{\partial \theta}.$$

But from above such a choice corresponds to the natural interpretation of the parameter vector since it leads not only to a consistent M-estimator but also to an M-estimator whose asymptotic variance covariance matrix is as small as possible.

For instance, when $f(y;\theta)$ is the normal density $N(\theta,1)$, then we have $\partial \log f(y;\theta)/\partial\theta = y-\theta$. Hence the estimating equation $E_\theta g(Y;\theta) = E_\theta(Y-\theta) = 0$, which leads to interpreting θ as the mean, is preferable to the estimating equation $E_\theta(\mathbb{1}_{Y>\theta} - 1/2) = 0$, which leads to interpreting θ as the median.

Example 23.2:

(i) Consider a random sample $(Y_1,\ldots,Y_n)$ drawn from a common probability distribution belonging to the exponential family with parameter $\theta \in \Theta \subset \mathbb{R}$. The probability density of one observation is of the form

$$f(y;\theta) = C(\theta) \exp \sum_{j=1}^{K} Q_j(h(\theta)) S_j(y),$$

where

$$h(\theta) = E_\theta S(Y) = (E_\theta S_1(Y),\ldots,E_\theta S_K(Y))'.$$

The log-likelihood function is

$$\log \ell(y;\theta) = n \log C(\theta) + n \sum_{j=1}^{K} Q_j(h(\theta)) \bar{S}_{jn},$$

where $\bar{S}_{jn} = \frac{1}{n}\sum_{i=1}^{n} S_j(Y_i)$. Hence the score is given by

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial \ell(y;\theta)}{\partial \theta} &= \sqrt{n}\frac{d\log C(\theta)}{d\theta} + \sqrt{n}\sum_{j=1}^{K}\frac{dh'(\theta)}{d\theta}\frac{dQ_j(h(\theta))}{dh}\bar{S}_{jn} \\ &= \sum_{j=1}^{K}\frac{dh'(\theta)}{d\theta}\frac{dQ_j(h(\theta))}{dh}\sqrt{n}(\bar{S}_{jn} - h_j(\theta)).\end{aligned}$$

(ii) To estimate the parameter θ, we can restrict ourselves to functions of the sufficient statistic $\bar{S}_n = (\bar{S}_{1n}, \ldots, \bar{S}_{Kn})'$. Here we shall consider such a function that is independent of the number n of observations and that is continuously differentiable. Let $T_n(Y) = \phi(\bar{S}_n)$.

The consistency condition for this estimator is

$$\phi(h(\theta)) = \theta.$$

It implies the first-order condition

$$\frac{d\phi(h(\theta))}{dh'}\frac{dh(\theta)}{d\theta} = 1.$$

Now, if we take a Taylor series expansion of the estimator, we obtain

$$\sqrt{n}(T_n(Y) - \theta) = \sqrt{n}(\phi(\bar{S}_n) - \phi(h(\theta))) \# \frac{d\phi(h(\theta))}{dh'}\sqrt{n}(\bar{S}_n - h(\theta)).$$

Hence the condition for first-order asymptotic efficiency is

$$\exists\, B(\theta) : \frac{d\phi(h(\theta))}{dh'} = B(\theta)\frac{dh'(\theta)}{d\theta}\frac{dQ'(h(\theta))}{dh}.$$

Right-multiplying both sides by the matrix $dh(\theta)/d\theta$ and taking into account the equalities on the derivatives that are implied by the consistency of T_n, we obtain

$$1 = B(\theta)\frac{dh'(\theta)}{d\theta}\frac{dQ'(h(\theta))}{dh}\frac{dh(\theta)}{d\theta}.$$

Therefore, under the restriction that it is consistent, the estimator $T_n = \phi(\bar{S}_n)$ is first-order asymptotically efficient if and only if the following condition on the partial derivatives of ϕ is satisfied

$$\frac{d\phi(h(\theta))}{dh'} = \frac{\dfrac{h'(\theta)}{d\theta}\dfrac{dQ'(h(\theta))}{dh}}{\dfrac{dh'(\theta)}{d\theta}\dfrac{dQ'(h(\theta))}{dh}\dfrac{dh(\theta)}{d\theta}}, \quad \forall\theta.$$

d) Superefficiency

When one considers estimators that do not satisfy Assumption 23.1, then the preceding results no longer hold. In particular, it is possible to construct estimators whose asymptotic variance covariance matrices are smaller than the asymptotic FDCR bound. Such estimators are said to be (asymptotically) *superefficient*.

Example 23.3: Constructing a superefficient estimator is straightforward. Let T_n be a first-order asymptotically efficient estimator of the scalar parameter θ. Assume that it satisfies Assumption 23.1. Define

$$\tilde{T}_n = \begin{cases} T_n, & \text{if } |T_n| \geq n^{-1/4}, \\ \alpha T_n, & \text{otherwise, where } \alpha \in [0,1). \end{cases}$$

We have

$$\sqrt{n}(\tilde{T}_n - \theta) = \sqrt{n}(T_n - \theta)\mathbb{1}_{|T_n| \geq n^{-1/4}} + \sqrt{n}(\alpha T_n - \theta)\mathbb{1}_{|T_n| < n^{-1/4}}.$$

If θ is nonzero, then $\sqrt{n}(\tilde{T}_n - \theta) \# \sqrt{n}(T_n - \theta)$. Hence

$$V_{as}(\sqrt{n}(\tilde{T}_n - \theta)) \# V_{as}(\sqrt{n}(T_n - \theta)) = \mathcal{I}_1(\theta)^{-1},$$

where $V_{as}(\cdot)$ denotes the asymptotic variance of the quantity in parentheses. On the other hand, if θ is equal to zero, then $\sqrt{n}(\tilde{T}_n - \theta) = \sqrt{n}\tilde{T}_n \# \alpha\sqrt{n}T_n$. Hence

$$V_{as}(\sqrt{n}\tilde{T}_n) = \alpha^2 \mathcal{I}_1(\theta)^{-1} < \mathcal{I}_1(\theta)^{-1}.$$

Note that, although asymptotically normal, $\sqrt{n}(\tilde{T}_n - \theta)$ does not converge uniformly in θ.

This example can be viewed as an artificial example. In fact, it can be proved that an analogous property holds for any superefficient estimator. Namely, one can show that the FDCR bound can be improved upon only over a set of parameter values that is negligible with respect to the Lebesgue measure. In the present example this set is $\{0\}$.

23.1.2 Higher-Order Expansions

a) The Principle

The asymptotic study of the preceding subsection consists first in considering consistent estimators of θ, i.e., estimators T_n satisfying

$$T_n = \theta + o(1),$$

where $o(1)$ denotes a quantity that converges to zero. Then a series expansion up to the next order is considered. In general, we have seen that the next term in the expansion is of order $1/\sqrt{n}$. As a matter of fact, for commonly used estimators we have

$$T_n = \theta + \frac{A_n}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

where A_n is a quantity that converges in distribution to a random variable A and where $o(1/\sqrt{n})$ denotes a quantity that converges to zero faster than $1/\sqrt{n}$.

This expansion allowed us to approximate the variance of T_n, namely

$$V_\theta \sqrt{n}(T_n - \theta) \approx V_\theta A,$$

i.e.

$$V_\theta T_n = \frac{V_\theta A}{n} + o\left(\frac{1}{n}\right).$$

Lastly, we sought estimators for which the approximated variances were the smallest. This gave us the so-called first-order asymptotically efficient estimators.

As we saw in previous chapters, however, there exist many first-order asymptotically efficient estimators. Examples are the maximum likelihood estimator, any estimator obtained after one iteration of the score algorithm starting from an initial consistent estimator, estimators obtained through various approximations of the maximum likelihood estimator based on quasi generalized least squares methods, etc. If we now want to choose among these estimators, it is natural to consider a higher Taylor series expansion. In the "regular" cases, the next terms in such an expansion correspond to powers of $1/\sqrt{n}$. It suffices to retain the first three terms in the expansion, which gives

$$T_n = \theta + \frac{A_n}{\sqrt{n}} + \frac{B_n}{n} + \frac{C_n}{n\sqrt{n}} + o\left(\frac{1}{n\sqrt{n}}\right), \tag{23.1}$$

where A_n, B_n, and C_n converge in distribution to A, B, and C, respectively. Intuitively such a higher-order expansion should allow us to approximate the mean of the estimator T_n up to the order $(1/n\sqrt{n})$ and the variance and the mean squared error up to the order $1/n^2$.

As an illustration, suppose that the first two moments of A_n, B_n, and C_n are independent of n and that the parameter θ is a scalar. Then, we have

$$ET_n = \theta + \frac{EA}{\sqrt{n}} + \frac{EB}{n} + \frac{EC}{n\sqrt{n}} + o\left(\frac{1}{n\sqrt{n}}\right),$$

and

$$VT_n = \frac{VA}{n} + \frac{2\text{cov}(A,B)}{n\sqrt{n}} + \frac{VB + 2\text{cov}(A,C)}{n^2} + o\left(\frac{1}{n^2}\right).$$

The reason for considering the term of order $1/n\sqrt{n}$ in the expansion (23.1) is that $\text{cov}(A,B) = 0$ when the estimator T_n is appropriately chosen. Then the first term in the variance that is nonzero after the term of order $1/n$ is the term of order $1/n^2$.

By analogy to the first-order efficiency, it is natural to define a preorder on the set of considered estimators based upon their second-order approximations. Specifically, the *second-order approximated quadratic risk* is

$$E_2(T_n - \theta)(T_n - \theta)' = \frac{D_1}{n} + \frac{D_2}{n\sqrt{n}} + \frac{D_3}{n^2},$$

where E_2 means that only the first terms up to order $1/n^2$ in the expansion of the variance are taken into account in the assessment of the quadratic risk. Then an estimator T_{1n} is said to be *second-order preferred* to another estimator T_{2n} if

$$E_2(T_{1n} - \theta)(T_{1n} - \theta)' \preceq E_2(T_{2n} - \theta)(T_{2n} - \theta)', \quad \forall\theta,$$

where $\preceq$ denotes the usual preorder over symmetric positive semidefinite matrices, as before.

As the next example shows, however, it is easy to see that there does not exist an estimator that is optimal in this sense.

Example 23.4: Suppose that T_n^* is an optimal estimator of θ in the preceding sense. Define the modified estimator

$$T_n^{**} = T_n^*\left(1 - \frac{1}{n}\right) + \frac{\theta_0}{n}.$$

When $\theta = \theta_0$, the second-order approximated risk of T_n^{**} is

$$\begin{aligned} E_2(T_n^{**} - \theta_0)(T_n^{**} - \theta_0)' &= \left(1 - \frac{1}{n}\right)^2 E_2(T_n^* - \theta_0)(T_n^* - \theta_0)' + o\left(\frac{1}{n^2}\right) \\ &\preceq E_2(T_n^* - \theta_0)(T_n^* - \theta_0)'. \end{aligned}$$

This contradicts the second-order optimality of T_n^*.

In fact such a difficulty was already encountered in the theory of optimal decisions. A possible solution then consisted in restricting estimators to be unbiased. A similar condition can also be introduced here.

Definition 23.5: *An estimator T_n satisfying the expansion (23.1) is second-order asymptotically unbiased or asymptotically unbiased at the order $1/n$ if*

$$ET_n = \theta + o\left(\frac{1}{n}\right).$$

b) The Second-Order Asymptotic Bias

In many cases the second-order asymptotic bias or the asymptotic bias of order $1/n$ can be easily determined from the terms A_n and B_n in the expansion of T_n. Three examples given next illustrate the method.

Example 23.5: Consider the case of an M-estimator $\tilde{\theta}_n$. To simplify the notation, we assume that the parameter θ is a scalar. The second-order Taylor series expansion of the first-order condition

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \psi(Y_i; \tilde{\theta}_n)}{\partial \theta}$$

gives

$$\begin{aligned} 0 \;=\; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \psi(Y_i;\theta)}{\partial \theta} + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \psi(Y_i;\theta)}{\partial \theta^2} \sqrt{n}(\tilde{\theta}_n - \theta) \\ & + \frac{1}{2\sqrt{n}} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3 \psi(Y_i;\theta)}{\partial \theta^3} (\sqrt{n}(\tilde{\theta}_n - \theta))^2 + o\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Hence we obtain

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \psi(Y_i;\theta)}{\partial \theta} + E_\theta \left(\frac{\partial^2 \psi(Y;\theta)}{\partial \theta^2} \right) \sqrt{n}(\tilde{\theta}_n - \theta) \\ & + \frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\partial^2 \psi(Y_i;\theta)}{\partial \theta^2} - E_\theta \left(\frac{\partial^2 \psi(Y;\theta)}{\partial \theta^2} \right) \right) \sqrt{n}(\tilde{\theta}_n - \theta) \\ & + \frac{1}{2\sqrt{n}} E_\theta \left(\frac{\partial^3 \psi(Y;\theta)}{\partial \theta^3} \right) (\sqrt{n}(\tilde{\theta}_n - \theta))^2 + o\left(\frac{1}{\sqrt{n}}\right) = 0. \end{aligned}$$

Replacing $\sqrt{n}(\tilde{\theta}_n - \theta)$ by the asymptotic equivalent quantity

$$\left(E_\theta \left(-\frac{\partial^2 \psi(Y;\theta)}{\partial \theta^2} \right) \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \psi(Y_i;\theta)}{\partial \theta}$$

in the last two terms, and letting

$$\mathcal{J} = E_\theta\left(-\frac{\partial^2\psi(Y;\theta)}{\partial\theta^2}\right),$$

we obtain

$$\begin{aligned}\tilde{\theta}_n = \theta &+ \frac{1}{\sqrt{n}}\mathcal{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\psi(Y_i;\theta)}{\partial\theta} \\ &+ \frac{1}{n}\left(\mathcal{J}^{-2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial^2\psi(Y_i;\theta)}{\partial\theta^2} - E_\theta\left(\frac{\partial^2\psi(Y;\theta)}{\partial\theta^2}\right)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\psi(Y_i;\theta)}{\partial\theta}\right. \\ &+ \left.\frac{1}{2}\mathcal{J}^{-3}E_\theta\left(\frac{\partial^3\psi(Y;\theta)}{\partial\theta^3}\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\psi(Y_i;\theta_n)}{\partial\theta}\right)^2\right) + o\left(\frac{1}{n}\right).\end{aligned}$$

Now, when the objective function is such that the M-estimator is consistent, then $E_\theta(\partial\psi(Y;\theta)/\partial\theta) = 0$. Hence the asymptotic bias of order $1/n$ is

$$\begin{aligned}E_2\tilde{\theta}_n - \theta \;=\; &\frac{1}{n}\left(\mathcal{J}^{-2}\mathrm{cov}_\theta\left(\frac{\partial^2\psi(Y;\theta)}{\partial\theta^2}, \frac{\partial\psi(Y;\theta)}{\partial\theta}\right)\right. \\ &\left.+\frac{1}{2}\mathcal{J}^{-3}E_\theta\left(\frac{\partial^3\psi(Y;\theta)}{\partial\theta^3}\right)V_\theta\left(\frac{\partial\psi(Y;\theta)}{\partial\theta}\right)\right).\end{aligned}$$

In particular, we note that, if the objective function is the log-likelihood function, i.e., if $\tilde{\theta}_n$ is the maximum likelihood estimator of θ, then the term of order $1/n$ is nonzero, in general. Therefore the maximum likelihood estimator is typically second-order asymptotically biased.

Example 23.6: Consider now Example 23.2 where the parametric family is the exponential family. The log-likelihood function is

$$\log\ell(y;\theta) = n\log C(\theta) + n\sum_{j=1}^{K}Q_j(h(\theta))\bar{S}_{jn}.$$

Consider an estimator of the form $T_n(Y) = \phi(\bar{S}_n)$. Taking a Taylor series expansion gives

$$\begin{aligned}\sqrt{n}(T_n(Y) - \theta) = &\frac{d\phi(h(\theta))}{dh'}\sqrt{n}(\bar{S}_n - h(\theta)) \\ &+\frac{1}{2\sqrt{n}}\sqrt{n}(\bar{S}_n - h(\theta))'\frac{d^2\phi(h(\theta))}{dhdh'}\sqrt{n}(\bar{S}_n - h(\theta)) + o\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Hence, we have

$$E_2 T_n(Y) = \theta + \frac{1}{n}\frac{1}{2}\,\mathrm{tr}\left(\frac{d^2\phi(h(\theta))}{dhdh'} V_\theta S(Y)\right) + o\left(\frac{1}{n}\right).$$

Example 23.7: Lastly, consider the linear model $Y_i = aX_i + u_i$, where X_i is a scalar random variable and (u_i, X_i), $i = 1, \ldots, n$, are independently and identically distributed with $Eu_iX_i = 0$. The OLS estimator is

$$\hat{a}_n = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2} = a + \frac{\sum_{i=1}^n X_i u_i}{\sum_{i=1}^n X_i^2}.$$

The OLS estimator converges to $a + E(Xu)/EX^2 = a$. Moreover, it is asymptotically normally distributed.

To determine the second-order asymptotic bias, we take a Taylor series expansion as in the previous example. This gives

$$\begin{aligned}
\sqrt{n}(\hat{a}_n - a) &= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i u_i}{EX^2\left(1 + \frac{1}{n}\sum_{i=1}^n \frac{(X_i^2 - EX^2)}{EX^2}\right)} \\
&= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i u_i}{EX^2}\left(1 - \frac{1}{n}\sum_{i=1}^n \frac{(X_i^2 - EX^2)}{EX^2}\right) + o\left(\frac{1}{\sqrt{n}}\right) \\
&= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i u_i}{EX^2} \\
&\quad \frac{1}{\sqrt{n}}\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i u_i \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i^2 - EX^2)}{(EX^2)^2} + o\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

It follows that

$$E_2\hat{a}_n = a - \frac{1}{n}\frac{E(X^3u)}{(EX^2)^2} + o\left(\frac{1}{n}\right).$$

Therefore the second-order asymptotic bias is equal to zero if and only if $E(X^3u) = 0$.

c) Eliminating the Second-Order Asymptotic Bias

Suppose that we restrict estimators to be second-order asymptotically unbiased. Then it is easy to see that one can eliminate the second-order

asymptotic bias of any consistent estimator satisfying an expansion of the form 23.1.

Property 23.5: *Let T_n be a consistent estimator of θ with a second-order asymptotic bias satisfying*

$$E_2 T_n = \theta + \frac{b(\theta)}{n},$$

where b is a continuous function. Then the estimator $T_n^* = T_n - b(T_n)/n$ is second-order asymptotically unbiased.

PROOF: It suffices to take a Taylor series expansion of $b(T_n)/n$. The leading term is $b(\theta)/n$. □

In particular, the second-order asymptotic bias of the maximum likelihood estimator can be eliminated in this manner.

d) The Jackknife Principle

The preceding method requires a precise knowledge of the bias function. The Jackknife principle, which is presented next, is frequently easier to apply. It relies explicitly on the form of the sequence of estimators.

Let $T_n = T_n(Y_1, \ldots, Y_n)$ be the estimator of θ obtained from n observations. Then other estimators can be obtained by deleting one observation each time. Namely, we can consider the estimators

$$T_n^i = T_{n-1}(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n), \quad i = 1, \ldots, n.$$

Property 23.6: *Let T_n be an estimator of θ with asymptotic mean satisfying*

$$E_\theta T_n = \theta + \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + o\left(\frac{1}{n^2}\right).$$

Then, the Jackknife estimator based on T_n defined by

$$J(T_n) = nT_n - \frac{n-1}{n} \sum_{i=1}^{n} T_{n-1}^i$$

satisfies

$$E_\theta J(T_n) = \theta + o\left(\frac{1}{n}\right).$$

PROOF: We have

$$\begin{aligned} E_\theta J(T_n) &= n\left(\theta + \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + o\left(\frac{1}{n^2}\right)\right) \\ &\quad -(n-1)\left(\theta + \frac{b_1(\theta)}{n-1} + \frac{b_2(\theta)}{(n-1)^2} + o\left(\frac{1}{n^2}\right)\right) \\ &= \theta - \frac{b_2(\theta)}{n(n-1)} + o\left(\frac{1}{n^2}\right) = \theta + o\left(\frac{1}{n}\right). \end{aligned}$$

□

Similarly, introducing estimators obtained by deleting two observations each time, it is possible to reduce the bias of order higher than $1/n$, i.e., to eliminate the term of order $1/n^2$.

e) Second-Order Asymptotic Efficiency

We now consider estimators that are second-order asymptotically unbiased or asymptotically unbiased of order $1/n$, i.e., such that $E_2 T_n = \theta + o(1/n)$. Then it is possible to derive a bound of order $1/n^2$ for the approximated risk $E_2(T_n - \theta)(T_n - \theta)'$. This result, which is likely to hold in the general case, was proved by Ghosh and Subramanyam (1974) in the case of a scalar parameter in exponential models. This result is formally stated in the next property.

Property 23.7:

(i) The maximum likelihood estimator corrected for the second-order bias

$$\hat{\theta}_n^* = \hat{\theta}_n - \frac{b(\hat{\theta}_n)}{n}$$

satisfies

$$E_2(\hat{\theta}_n^* - \theta)^2 = \frac{\mathcal{I}_1^{-1}}{n} + \frac{D}{n^2} + o\left(\frac{1}{n^2}\right).$$

(ii) The approximated risk of the estimator $\hat{\theta}_n^*$ is smaller or equal to the approximated risk of any other asymptotically unbiased estimator of order $1/n$.

By analogy to the definition of first-order asymptotic efficiency, we can give the following definition.

Definition 23.6: *A second-order asymptotically unbiased estimator is second-order asymptotically efficient if its approximated risk is equal to that of $\hat{\theta}_n^*$.*

f) Newton–Raphson Algorithm and Asymptotic Efficiency

Various algorithms are available to determine numerically the maximum likelihood estimator (see Chapter 13). We shall consider the Newton–Raphson algorithm, which is based on the following iteration formula

$$\begin{aligned}\tilde{\theta}_{p+1} &= \tilde{\theta}_p - \left(\sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i;\tilde{\theta}_p)}{\partial\theta^2}\right)^{-1} \sum_{i=1}^{n} \frac{\partial \log f(Y_i;\tilde{\theta}_p)}{\partial\theta} \\ &= \tilde{\theta}_p - \left(\frac{\partial^2 \log \ell(Y;\tilde{\theta}_p)}{\partial\theta^2}\right)^{-1} \frac{\partial \log \ell(Y;\tilde{\theta}_p)}{\partial\theta}, \quad \forall p \in I\!N,\end{aligned}$$

where we have assumed that the parameter θ is scalar to simplify.

Suppose now that the inital value $\tilde{\theta}_0$ corresponds to an estimator θ_n^* that differs from the maximum likelihood estimator $\hat{\theta}_n$ by a term of order $1/\sqrt{n}$. That is

$$\theta_n^* = \hat{\theta}_n + \frac{A_n}{\sqrt{n}},$$

where A_n converges in distribution. Then, taking a Taylor series expansion around the true parameter value θ_0, we obtain

$$\tilde{\theta}_{p+1} - \hat{\theta}_n = \tilde{\theta}_p - \hat{\theta}_n$$

$$- \frac{\frac{\partial^2 \log \ell(Y;\theta_0)}{\partial\theta^2}(\tilde{\theta}_p - \hat{\theta}_n) + \frac{1}{2}\frac{\partial^3 \log \ell(Y;\theta_0)}{\partial\theta^3}((\tilde{\theta}_p - \theta_0)^2 - (\hat{\theta}_n - \theta_0)^2) + \ldots}{\frac{\partial^2 \log \ell(Y;\theta_0)}{\partial\theta^2} + \frac{\partial^3 \log \ell(Y;\theta_0)}{\partial\theta^3}(\tilde{\theta}_p - \theta_0) + \ldots},$$

which gives

$$\begin{aligned}\tilde{\theta}_{p+1} - \hat{\theta}_n &= \tilde{\theta}_p - \hat{\theta}_n - (\tilde{\theta}_p - \hat{\theta}_n) \\ &\times \left(1 + \frac{1}{2}\frac{\frac{\partial^3 \log \ell(Y;\theta_0)}{\partial\theta^3}}{\frac{\partial^2 \log \ell(Y;\theta_0)}{\partial\theta^2}}(\tilde{\theta}_p - \theta_0 + \hat{\theta}_n - \theta_0 - 2\tilde{\theta}_p + 2\theta_0) + \ldots\right),\end{aligned}$$

or

$$\tilde{\theta}_{p+1} - \hat{\theta}_n \sim \frac{1}{2}\frac{E\left(\frac{\partial^3 \log f(Y;\theta_0)}{\partial\theta^3}\right)}{\left(\frac{\partial^2 \log f(Y;\theta_0)}{\partial\theta^2}\right)}(\tilde{\theta}_p - \hat{\theta}_n)^2.$$

Because $\tilde{\theta}_0 - \hat{\theta}_n = \theta_n^* - \hat{\theta}_n$ is of order $1/\sqrt{n}$, it follows from the preceding equation that $\tilde{\theta}_1 - \hat{\theta}_n$ is of order $1/n$, $\tilde{\theta}_2 - \hat{\theta}_n$ is of order $1/n^2$, etc. Hence we obtain the following property.

Property 23.8: *Let θ_n^* be an estimator of the parameter θ. Suppose that the difference between θ_n^* and the maximum likelihood estimator is of order $1/\sqrt{n}$.*

(i) The estimator $\tilde{\theta}_1$ obtained after one Newton–Raphson iteration is equivalent at the order $1/\sqrt{n}$ to the maximum likelihood estimator, i.e.

$$\sqrt{n}(\tilde{\theta}_1 - \theta) = \sqrt{n}(\hat{\theta}_n - \theta) + o(1).$$

(ii) The estimator $\tilde{\theta}_2$ obtained after two Newton–Raphson iterations is such that its expansion up to the order $1/(n\sqrt{n})$ is identical to that of the maximum likelihood estimator. Thus, after correcting for its second-order bias, this estimator provides an estimator that is second-order asymptotically efficient.

23.1.3 Edgeworth Expansion

When evaluating the second-order asymptotic bias in Examples 23.5, 23.6, and 23.7, we applied in fact different methods for expanding the statistic so as to determine the random variables A_n and B_n appearing in the decomposition 23.1. Some of these methods can be generalized to determine the additional term C_n and, hence, can be used to determine whether the estimator is second-order asymptotically efficient. These derivations are frequently relatively complex. A general approach, which is somewhat different, can be invoked when the main object of interest is the decomposition of the probability distribution of the statistic and not the random variables A_n, B_n, and C_n themselves. Such an alternative approach is especially useful when the estimator is an empirical mean or a function of empirical means. To simplify the exposition, we shall consider the first situation only.

We consider a sequence of independently and identically random variables $Y_1, \ldots, Y_n, \ldots$, with zero means, variances equal to 1, and moments up to the order 4. Then we can expand the logarithm of the characteristic function of the probability distribution of Y up to the order 4. This gives

$$\log \psi(t) = -\frac{1}{2}t^2 + \frac{1}{6}k_3(it)^3 + \frac{1}{24}k_4(it)^4 + o(t^4),$$

where $\Psi(t) = E(\exp itY)$, $k_3 = EX^3$ and $k_4 = EX^4 - 3$. Hence the

characteristic function of the normalized sum $T_n = (1/\sqrt{n}) \sum_{i=1}^{n} Y_i$ satisfies

$$\begin{aligned}\Psi_n(t) &= \Psi\left(\frac{t}{\sqrt{n}}\right)^n \\ &= \exp\left(-\frac{1}{2}t^2 + \frac{1}{6\sqrt{n}}k_3(it)^3 + \frac{1}{24n}k_4(it)^4\right) + o\left(\frac{1}{n}\right) \\ &= \exp\left(-\frac{1}{2}t^2\right)\left(1+\frac{1}{6\sqrt{n}}k_3(it)^3+\frac{1}{72n}(3k_4(it)^4+k_3^2(it)^6)\right)+o\left(\frac{1}{n}\right).\end{aligned}$$

Thus we obtain an expansion of the characteristic function of T_n in powers of $1/\sqrt{n}$. It follows that we can obtain an expansion of the probability density of T_n by inverting the Fourier transform. If $f_n(t)$ denotes the probability density of T_n, we obtain

$$f_n(t) = \phi(t) - \frac{1}{6\sqrt{n}}k_3\phi^{(3)}(t) + \frac{1}{24n}k_4\phi^{(4)}(t) + \frac{1}{72n}k_3^2\phi^{(6)}(t) + o\left(\frac{1}{n}\right),$$

where ϕ denotes the probability density of the standard normal distribution. Because the derivatives of the probability density ϕ are of the form $\phi(t)H(t)$, namely $\phi^{(r)} = (-1)^r\phi(t)H_r(t)$, where H_r is a polynomial called a *Hermite polynomial*, we have

$$f_n(t) = \phi(t)\left(1 - \frac{1}{6\sqrt{n}}k_3H_3(t) + \frac{1}{24n}k_4H_4(t) + \frac{1}{72n}k_3^2H_6(t)\right)+o\left(\frac{1}{n}\right).$$

Then, integrating with respect to t, we obtain the so-called Edgeworth approximation of the cumulative distribution function, namely

$$F_n(t) = \Phi(t) - \phi(t)\left(\frac{k_3H_2(t)}{6\sqrt{n}} + \frac{k_4H_3(t)}{24n} + \frac{k_3^2H_5(t)}{72n}\right) + o\left(\frac{1}{n}\right), \tag{23.2}$$

where Φ is the cumulative distribution function of the standard normal. The Hermite polynomials are given by

$$\begin{aligned}H_2(t) &= t^2 - 1, \\ H_3(t) &= t^3 - 3t, \\ H_4(t) &= t^4 - 6t^2 + 3, \\ H_5(t) &= t^5 - 10t^3 + 15t.\end{aligned}$$

An alternative approximated form, which gives identical leading terms up to the order $1/n$ and which is frequently more convenient, is

$$F_n(t) = \Phi\left(t - \frac{k_3(t^2-1)}{6\sqrt{n}} + \frac{3k_4(3t-t^3) + 2k_3^2(4t^3-7t)}{72n}\right) + o\left(\frac{1}{n}\right). \tag{23.3}$$

Hence we obtain an approximation formula of the form

$$P(T_n \leq t) \# \Phi\left(t + \frac{C_1(t)}{\sqrt{n}} + \frac{C_2(t)}{n}\right).$$

When n is large, the function $t \to t + C_1(t)/\sqrt{n} + C_2(t)/n$ is increasing in t. Therefore an equivalent approximation is

$$P\left(T_n + \frac{C_1(T_n)}{\sqrt{n}} + \frac{C_2(T_n)}{n} \leq t\right) \# \Phi(t).$$

23.2 Asymptotic Comparison of Tests

The theory for the asymptotic comparison of tests is somewhat similar to that of estimators. Specifically, one first introduces the concept of consistent tests. Because there exist, in general, many consistent tests, then one seeks to determine whether there exists an optimal test among those consistent tests. Unfortunately, in the case of tests there does not exist, in general, a consistent test that is "asymptotically uniformly most powerful." In view of such a difficulty, statisticians have proposed various methods for comparing power functions *locally*. These methods are presented in this section.

23.2.1 Statement of the Problem and Various Solutions

a) An Example of a Consistent Test

We consider a random sample $Y_1, \ldots, Y_n$, where the random variables are distributed $N(\theta, 1)$. We are interested in testing the null hypothesis H_0: $\theta \leq 0$ against the alternative H_1: $\theta > 0$. A commonly used test at level α is the test with critical region

$$W(n, \alpha) = \{\sqrt{n}\bar{Y}_n > \Phi^{-1}(1-\alpha)\},$$

where Φ denotes the cumulative distribution function of the standard normal, as before. The power function of such a test is

$$\begin{aligned}\gamma(n,\alpha,\theta) &= P_\theta(W(n,\alpha)) \\ &= P(\sqrt{n}(\bar{Y}_n - \theta) > \Phi^{-1}(1-\alpha) - \sqrt{n}\theta),\end{aligned}$$

where $\theta > 0$. This gives

$$\gamma(n,\alpha,\theta) = 1 - \Phi(\Phi^{-1}(1-\alpha) - \sqrt{n}\theta). \tag{23.4}$$

Thus the power function associated with the test with critical region $W(n,\alpha)$ depends on three factors, which are the number n of observations, the level α of the test, and the value of the mean parameter θ. In this example it is easy to see that the power function is increasing in n, increasing in α, and increasing in θ. Moreover, the power function has the following asymptotic properties:

$$\begin{aligned}\text{for } \alpha = 0 \quad &: \quad \gamma(n,0,\theta) = 0, \\ \text{for } \theta = 0 \quad &: \quad \gamma(n,\alpha,0) = \alpha, \\ \text{for large } n \quad &: \quad \lim_{n\to\infty} \gamma(n,\alpha,\theta) = 1, \text{ if } \alpha > 0 \text{ and } \theta > 0.\end{aligned}$$

The corresponding family of tests is consistent since the power converges to 1 as the number n of observations increases to infinity, when $\alpha > 0$ and $\theta > 0$. However, because the limit function is discontinuous when $\alpha = 0$ or $\theta = 0$, the convergence cannot be uniform.

Figure 23.1 illustrates how the power function changes for a given level α, as n increases to infinity.

b) Exact and Approximate Methods

We now consider two families of tests associated with the critical regions $W^1(n,\alpha)$ and $W^2(n,\alpha)$. If, as in the preceding example, the power functions have some simple explicit forms, then a direct comparison can be carried out asymptotically or in finite samples. Unfortunately, this is seldom the case in practice. For, in general,

(i) either the power function $\gamma(n,\alpha,\theta)$, $\forall n,\alpha,\theta$, can be determined, but this function is typically difficult to use because it introduces gamma functions, series expansions, etc.,

(ii) or the power function is unknown or cannot be determined explicitly.

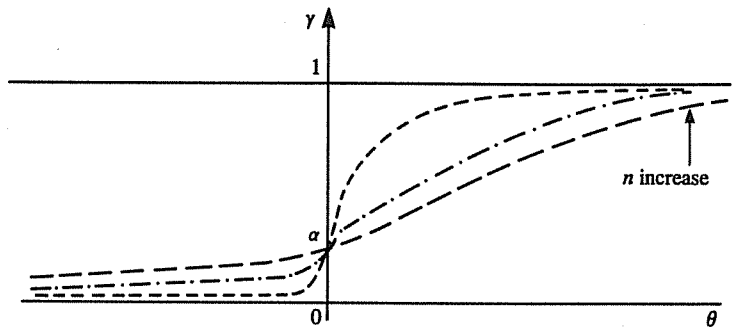


Figure 23.1: Power Functions

Making a distinction between the two cases (i) and (ii) seems important. For, in the former case, we can frequently construct a test with critical region $W(n, \alpha)$ and exact level α, i.e., a test satisfying

$$\sup_{\theta \in \Theta_0} P_\theta(W(n, \alpha)) = \alpha.$$

Then the approach is said to be *exact.*

In the latter case we often know only the asymptotic distribution of the test statistic under the null hypothesis. As a result, we can only construct a critical region whose level is α when n increases to infinity. That is, the corresponding family of critical regions satisfies

$$\sup_{\theta \in \Theta_0} \lim_{n \to \infty} P_\theta(W(n, \alpha)) = \alpha.$$

This approach, which is *asymptotic*, was used in the preceding chapters when studying the log-likelihood ratio test, the Wald test, the score test, the Lagrange multipliers test, etc. The approach is approximate in the sense that

$$\sup_{\theta \in \Theta_0} P_\theta(W(n, \alpha)) \neq \alpha$$

in finite samples.

c) Direct Methods

We consider again two families of tests defined by the critical regions $W_1(n, \alpha)$ and $W_2(n, \alpha)$ with exact or approximate level α.

A first idea for comparing these two families of tests consists in studying the ratio of the power functions $\gamma^1(n, \alpha, \theta)/\gamma^2(n, \alpha, \theta)$. At first sight, this ratio can be viewed as being analogous to the risk ratio in the case of estimators. Unfortunately, this idea is not of much use. For, if we require tests to be consistent, which is a natural requirement, then we have

$$\lim_{n \to \infty} \frac{\gamma^1(n, \alpha, \theta)}{\gamma^2(n, \alpha, \theta)} = 1, \quad \forall \theta \in \Theta.$$

Hence the ratio of power functions cannot be used to discriminate asymptotically between two testing procedures.

Two alternative approaches can be considered. The first approach consists in determining which of the two power functions converges faster to 1. This approach can be followed in some special situations. In such cases expanding $\log \gamma(n, \alpha, \theta)$ is frequently preferred to expanding $\gamma(n, \alpha, \theta)$. This approach, however, has not led to a general theory.

The second approach consists in building a framework in which the limit of the power function is no longer equal to one when the number of observations increases to infinity. Because the convergence of the power function is not uniform, it suffices to consider situations that become closer to those that raise difficulties. In the example of Section 23.2.1-a), these difficulties arise when $\theta = 0$ or $\alpha = 0$.

For instance, in this example we can seek a sequence of parameter values $\theta_n \in \Theta_1$ such that $\lim_{n \to \infty} \theta_n = 0$ and such that $\lim_{n \to \infty} \gamma^1(n, \alpha, \theta_n)$ and $\lim_{n \to \infty} \gamma^2(n, \alpha, \theta_n)$ exist and are different from 0 and 1. Such a sequence θ_n defines what is called a *sequence of local alternatives.* Then, if the ratio of the limits of the power functions satisfies

$$\frac{\lim_{n \to \infty} \gamma^1(n, \alpha, \theta_n)}{\lim_{n \to \infty} \gamma^2(n, \alpha, \theta_n)} < 1,$$

the second family of tests is preferred to the first family of tests. Note, however, that the limit ratio depends on the value α and on the chosen sequence of local alternatives. Hence, even if the above equality holds, there is no reason a priori that the family defined by W^2 is uniformly preferred to the family defined by W_1.

Alternatively, we can let the level α_n converge to zero as the number n of observations increases to infinity. Then we can seek to determine the appropriate rate of convergence of α_n for which $\gamma(n, \alpha_n, \theta)$ converges to a limit that is different from 0 and 1.

Example 23.8: Consider the example of Section 23.2.1-a), where a random sample is drawn from the normal distribution $N(\theta, 1)$. The

power function of the chosen critical region is

$$\gamma(n, \alpha, \theta) = 1 - \Phi(\Phi^{-1}(1-\alpha) - \sqrt{n}\theta).$$

(i) Suppose that the level α remains constant and let θ converge to zero. The limit of the power function differs from 0 and 1 only if θ_n is of the form $\delta/\sqrt{n}$, for some $\delta > 0$. Then we have

$$\lim_{n\to\infty} \gamma\left(n, \alpha, \frac{\delta}{\sqrt{n}}\right) = 1 - \Phi(\Phi^{-1}(1-\alpha) - \delta).$$

As noted above, this limit depends on the chosen level α and the "rate of convergence" of the sequence of local alternatives through δ.

(ii) Alternatively, we can consider a "symmetric" approach, where θ remains fixed and the level α_n converges to zero. To compensate for the fact that $\sqrt{n}\theta$ increases to infinity, it is necessary to choose α_n such that

$$\Phi^{-1}(1-\alpha_n) \sim \sqrt{n}\theta,$$

i.e., such that

$$1 - \alpha_n \sim \Phi(\sqrt{n}\theta).$$

But, from Property B.33, we know that

$$\Phi(x) \sim 1 - \frac{\phi(x)}{x}$$

when x is large, where ϕ is the probability density of the $N(0,1)$ distribution. Hence

$$1 - \alpha_n \sim 1 - (\sqrt{2\pi n}\theta)^{-1} \exp\left(-\frac{n\theta^2}{2}\right),$$

i.e.

$$\alpha_n \sim (\sqrt{2\pi n}\theta)^{-1} \exp\left(-\frac{n\theta^2}{2}\right).$$

Therefore the sequence of levels α_n ensuring the convergence of the power function to a number different from 0 and 1 depends on the value of the parameter θ.

d) Dual Methods

The preceding methods are not always easy to apply. As a result, they have been frequently replaced by an alternative approach, which is based

on the sample size that is necessary to ensure a given level of power. Such a minimal sample size depends on the level α, the critical level γ corresponding to the chosen power level, and the parameter θ. This minimal sample size is formally defined as

$$N(\alpha, \gamma, \theta) = \inf\{n : \gamma(n, \alpha, \theta) \geq \gamma\}. \tag{23.5}$$

Example 23.9: Consider again the power function of the test of the null hypothesis H_0: $\theta \leq 0$ using a Gaussian random sample. We have

$$\begin{aligned} \gamma(n, \alpha, \theta) \geq \gamma &\Leftrightarrow 1 - \Phi(\Phi^{-1}(1-\alpha) - \sqrt{n}\theta) \geq \gamma \\ &\Leftrightarrow \Phi^{-1}(1-\gamma) \geq \Phi^{-1}(1-\alpha) - \sqrt{n}\theta \\ &\Leftrightarrow n \geq \left(\frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\gamma)}{\theta}\right)^2. \end{aligned}$$

Thus, the minimal sample size is

$$N(\alpha, \gamma, \theta) = \left(\frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\gamma)}{\theta}\right)^2.$$

Now, to study local properties, these minimal sample sizes are compared when the level α converges to zero or when the parameter θ converges to zero from above. Note that in either case the minimal sample size converges to infinity. For instance, the sample size must increase to infinity so as to ensure a zero level. Similarly, the sample size must increase to infinity so as to distinguish between two parameter values that are infinitely close to each other.

The dual approach based on the convergence of θ_n to a point in the null hypothesis (to zero in the preceding example) is due to Pitman (1979). The dual approach based on the convergence of the level α_n to zero was developed by Bahadur (1960, 1967).

Table 23.1 summarizes the various methods that can be used to compare competing testing procedures asymptotically.

23.2.2 Concepts

a) Tests, Power, Levels

Hereafter, we consider testing a hypothesis in a parametric model. The null hypothesis is H_0: $\theta \in \Theta_0$ and the alternative hypothesis is H_1: $\theta \in \Theta_1$.

Table 23.1

Comparison based on → using ↓	the power function (Direct Approach)	the minimal number of observations (Dual Approach)
Local alternatives	frequently used in econometrics	Pitman's approach
Risk of Type I error converging to zero	seldom used	Bahadur's approach

The tests that we shall compare are assumed to be based upon some univariate statistics T_n. The alternative hypothesis is accepted when the value of the test statistic is sufficiently large. It is rejected otherwise. Hence each test is a nonrandomized test with a critical region of the form

$$W(n, \alpha) = \{T_n \geq t(n, \alpha)\},$$

where the critical value $t(n, \alpha)$ typically depends on the number n of observations and the chosen level α of the test.

It remains to explain how the critical value $t(n, \alpha)$ is determined. To do so, we distinguish between exact and approximate methods.

(i) EXACT METHODS

The critical value is obtained by using the exact probability distribution of the test statistic T_n under the null hypothesis. Let

$$\Lambda_n(t) = \sup_{\theta \in \Theta_0} P_\theta(T_n \geq t), \tag{23.6}$$

where Λ_n is the upper bound of the survival function of T_n under the null hypothesis. Such a function is decreasing in t. To simplify the exposition, it is assumed that this function is continuous and strictly decreasing. Hence it is invertible.

It follows that the exact critical region is given by

$$W(n, \alpha) = \{T_n \geq \Lambda_n^{-1}(\alpha)\} = \{\Lambda_n(T_n) \leq \alpha\}. \tag{23.7}$$

Thus the power function is

$$\gamma(n, \alpha, \theta) = P_\theta(W(n, \alpha)) = P_\theta(\Lambda_n(T_n) \leq \alpha), \quad \theta \in \Theta_1. \tag{23.8}$$

Two concepts are closely related to the concept of power. The first concept is the *minimal number of observations* that is necessary to attain a given value γ of power. It is defined by

$$N(\alpha, \gamma, \theta) = \inf\{n : \gamma(n, \alpha, \theta) \geq \gamma\}. \tag{23.9}$$

The second concept is the *minimal level* α that is necessary to attain a given level γ of power. It is defined by

$$\alpha(n, \gamma, \theta) = \inf\{\alpha : \gamma(n, \alpha, \theta) \geq \gamma\}. \tag{23.10}$$

(ii) Approximate Methods

Suppose that the test statistic T_n converges in distribution as n increases to infinity. Then we can determine the critical value $t(n, \alpha)$ from the asymptotic distribution. Specifically, let Q_θ denote the asymptotic distribution of T_n when the parameter value is θ. Let

$$\Lambda(t) = \sup_{\theta \in \Theta_0} Q_\theta(T \geq t). \tag{23.11}$$

Then an approximate critical region is given by

$$W^a(n, \alpha) = \{T_n \geq \Lambda^{-1}(\alpha)\} = \{\Lambda(T_n) \leq \alpha\}. \tag{23.12}$$

The approximate power function is

$$\gamma^a(n, \alpha, \theta) = P_\theta(W^a(n, \alpha)) = P_\theta(\Lambda(T_n) \leq \alpha), \quad \forall \theta \in \Theta_1. \tag{23.13}$$

Then, as for exact methods, the concepts of minimal number of observations and minimal level can be defined by replacing the exact power function by the approximate power function.

Remark 23.1: In finite samples the approximate test is not of level α, in general. For we have

$$\sup_{\theta \in \Theta_0} P_\theta(W^a(n, \alpha)) = \sup_{\theta \in \Theta_0} P_\theta(T_n \geq \Lambda^{-1}(\alpha)) = \Lambda_n(\Lambda^{-1}(\alpha)).$$

Hence the latter quantity is not necessarily equal to α. In practice the function Λ_n is frequently unknown. Thus the actual level of the test in finite samples is unknown. In particular, the actual level of the test may be larger than α. This arises when $\Lambda_n(\Lambda^{-1}(\alpha)) > \alpha$.

b) Some General Results

Below we give various properties that relate the functions $\gamma(n, \alpha, \theta)$, $\alpha(n, \gamma, \theta)$, and $N(\alpha, \gamma, \theta)$. Though they are stated for exact methods, these properties also apply to approximate methods by replacing Λ_n, γ, α and N by Λ, γ^a, α^a, and N^a, respectively.

Property 23.9:

(i) The mapping $\alpha \rightarrow \gamma(n, \alpha, \theta)$ is increasing in α.

(ii) $N(\alpha, \gamma, \theta) = \inf\{n : \alpha(n, \gamma, \theta) \leq \alpha\}$.

PROOF:

(i) We have $\gamma(n, \alpha, \theta) = P_\theta(\Lambda_n(T_n) \leq \alpha)$. Then it suffices to note that the cumulative distribution function of $\Lambda_n(T_n)$ is increasing.

(ii) From (i) it follows that

$$\gamma(n, \alpha, \theta) \geq \gamma \Leftrightarrow \alpha(n, \gamma, \theta) \leq \alpha.$$

The desired result follows immediately. □

The next property relates the asymptotic properties of the test statistic to those of the function $\alpha(n, \gamma, \theta)$.

Property 23.10: *Let θ be a value in Θ_1. Let $c(\theta)$ be a strictly positive scalar. The following two statements are equivalent:*

(i) $-\frac{2}{n} \log \Lambda_n(T_n)$ *converges to* $c(\theta)$ *in* P_θ *probability.*

(ii) $\forall \gamma \in (0, 1)$, $\lim_{n \rightarrow \infty} -\frac{2}{n} \log \alpha(n, \gamma, \theta) = c(\theta)$.

SKETCH OF PROOF: By definition we know that

$$\alpha(n, \gamma, \theta) = \inf\{\alpha : P_\theta(\Lambda_n(T_n) \leq \alpha) \geq \gamma\},$$

which is equivalent to

$$-\frac{2}{n} \log \alpha(n, \gamma, \theta) = \sup\left\{\mu : P_\theta\left(-\frac{2}{n} \log \Lambda_n(T_n) \geq \mu\right) \geq \gamma\right\}.$$

The desired result follows immediately from this equality. For instance, to show that (i) implies (ii), we have that $-(2/n) \log \Lambda_n(T_n)$ converges to $c(\theta)$. Hence

$$\text{if } \mu < c(\theta) : \lim_{n \rightarrow \infty} P_\theta\left(-\frac{2}{n} \log \Lambda_n(T_n) \geq \mu\right) = 1,$$

$$\text{if } \mu > c(\theta) : \lim_{n \rightarrow \infty} P_\theta\left(-\frac{2}{n} \log \Lambda_n(T_n) \geq \mu\right) = 0.$$

Therefore

$$\lim_{n\to\infty} -\frac{2}{n}\log\alpha(n,\gamma,\theta) = c(\theta), \quad \forall\gamma \in (0,1),$$

as desired. □

Property 23.11: *Suppose that $\alpha(n,\gamma,\theta)$ is a decreasing function of n. Let $c(\theta)$ be a strictly positive scalar. If $\lim_{n\to\infty} -(2/n)\log\alpha(n,\gamma,\theta) = c(\theta)$, then $N(\alpha,\gamma,\theta)$ is equivalent to $-2\log\alpha/c(\theta)$ as α converges to zero.*

SKETCH OF THE PROOF: From Property 23.9 it follows that

$$\begin{aligned} N(\alpha,\gamma,\theta) &= \inf\{n : \alpha(n,\gamma,\theta) \leq \alpha\} \\ &= \inf\left\{n : -\frac{2}{n}\log\alpha(n,\gamma,\theta) \geq -\frac{2\log\alpha}{n}\right\}. \end{aligned}$$

If α converges to zero, then $N(\alpha,\gamma,\theta)$ increases to infinity. Moreover, we have

$$N(\alpha,\gamma,\theta) \# \inf\left\{n : c(\theta) \geq -\frac{2\log\alpha}{n}\right\} = -\frac{2\log\alpha}{c(\theta)}.$$

□

23.2.3 Bahadur's Relative Efficiency

A first notion of efficiency was introduced by Bahadur. We follow below his presentation, which starts from the concept of *slope*. Then, we show the equivalence between Bahadur's approach and the approach based on the minimal number of observations discussed in the preceding section.

a) Slopes

When one carries a test, one typically chooses a risk of Type I error α. Then one determines the corresponding critical value. Next, this critical value is compared to the value of the test statistic. Within Bahadur's approach, however, there is no reason to fix a priori the risk of Type I error since this risk will become smaller as the number of observations increases. Instead, one computes the value of the test statistic and, then, one determines the level of the test for which the value of the test statistic is equal to the critical value. Thus, in the example of Section 23.2.1-a) the level α_n must satisfy

$$\Phi^{-1}(1-\alpha_n) = \sqrt{n}\bar{Y}_n.$$

Hence it is given by

$$\alpha_n = 1 - \Phi(\sqrt{n}\bar{Y}_n).$$

In the general case we know that the exact critical value is $\Lambda_n^{-1}(\alpha)$ while the approximate critical value is $\Lambda^{-1}(\alpha)$. Therefore the so-called *level* or *limit level* is

$$\begin{aligned} \alpha_n &= \Lambda_n(T_n) \quad \text{(exact method)}, \\ \alpha_n^a &= \Lambda(T_n) \quad \text{(approximate method)}. \end{aligned} \tag{23.14}$$

Then, according to the exact method, the basic idea is to prefer a test based on test statistic T_n^1 over another test based on the test statistic T_n^2 if, at least asymptotically, $\Lambda_n(T_n^1)$ is less than $\Lambda_n(T_n^2)$. Similarly, according to the approximate method, the first test is preferred to the second test if $\Lambda(T_n^1) < \Lambda(T_n^2)$ at least asymptotically.

This basic idea suggests to examine the asymptotic behavior of the variable $\Lambda_n(T_n)$. If we consider again the example of Section 23.2.1-a) where the random sample is drawn from a normal distribution and the test bears on the mean of this distribution, then we have $T_n = \sqrt{n}\bar{Y}_n \sim N(\sqrt{n}\theta, 1)$. Hence

$$\begin{aligned} \Lambda_n(t) = \Lambda(t) &= \sup_{\theta \leq 0} P_\theta(T_n \geq t) \\ &= \sup_{\theta \leq 0} P_\theta(T_n - \sqrt{n}\theta \geq t - \sqrt{n}\theta) \\ &= \sup_{\theta \leq 0} \left[1 - \Phi(t - \sqrt{n}\theta)\right] = 1 - \Phi(t). \end{aligned}$$

It follows that $\Lambda_n(T_n) = 1 - \Phi(T_n) = 1 - \Phi(\sqrt{n}\bar{Y}_n)$.

Under the alternative hypothesis, i.e. when $\theta > 0$, we have

$$\begin{aligned} \Lambda_n(T_n) &= 1 - \Phi(\sqrt{n}\bar{Y}_n) \\ &\# \quad 1 - \Phi(\sqrt{n}\theta) \\ &\# \quad (\sqrt{2\pi n}\theta)^{-1} \exp -\frac{n\theta^2}{2}. \end{aligned}$$

Hence

$$\log \Lambda_n(T_n) \ \# \ -\frac{n\theta^2}{2},$$

i.e.

$$-\frac{2}{n} \log \Lambda_n(T_n) \ \# \ \theta^2.$$

This example shows that it is desirable to consider a transformation of the limit level instead of the limit level itself. This motivates the following definitions, which are frequently used.

Definition 23.7:

(i) The variable $K_n = -\frac{2}{n}\log\Lambda_n(T_n)$ *is called the exact slope of the test in Bahadur's sense.*

(ii) The variable $K_n^a = -\frac{2}{n}\log\Lambda(T_n)$ *is called the approximate slope of the test.*

We now consider two sequences of tests with respective (exact) slopes

$$K_n^1 = -\frac{2}{n}\log\Lambda_n^1(T_n^1), \qquad K_n^2 = -\frac{2}{n}\log\Lambda_n^2(T_n^2).$$

Then, intuitively, the first sequence of tests is asymptotically preferred to the second sequence if K_n^2 is asymptotically smaller than K_n^1, when n is large. More precisely, we have the following definition.

Definition 23.8: *Consider two sequences of tests based on* T_n^1 *and* T_n^2 *with respective slopes* K_n^1 *and* K_n^2. *Suppose that* K_n^i *converges to a strictly positive scalar* $c^i(\theta)$ *in* P_θ*-probability as* n *increases to infinity, for* $i = 1, 2$ *and* $\theta \in \Theta_1$.

(i) The asymptotic relative efficiency (ARE) in Bahadur's sense is the quantity

$$e^B(T^1, T^2, \theta) = \frac{c^1(\theta)}{c^2(\theta)}.$$

(ii) The scalar $c^i(\theta)$ *is called the asymptotic slope of the test based on* T_n^i.

(iii) The test based on T^1 *is asymptotically preferred to the test based on* T^2 *in Bahadur's sense under the alternative* θ *if and only if*

$$e^B(T^1, T^2, \theta) > 1.$$

Of course, similar definitions hold when approximate slopes are considered.

b) Other Interpretations of Bahadur's ARE

From Property 23.10 we know that the convergence of the slope K_n^i to the limit $c^i(\theta)$ is equivalent to the convergence of $-(2/n)\log\alpha^i(n,\gamma,\theta)$ to this same limit. It follows immediately that another definition of asymptotic relative efficiency can be based on the following property.

Property 23.12: *We have*

$$e^B(T^1,T^2,\theta) = \lim_{n\to\infty} \frac{\log\alpha^1(n,\gamma,\theta)}{\log\alpha^2(n,\gamma,\theta)}.$$

Similarly, using Property 23.11, we obtain another characterization of Bahadur's ARE.

Property 23.13: *We have*

$$e^B(T^1,T^2,\theta) = \lim_{\alpha\to 0} \frac{N_2(\alpha,\gamma,\theta)}{N_1(\alpha,\gamma,\theta)}.$$

Thus the test T^1 is preferred to the test T^2 in Bahadur's sense if, asymptotically, the risk of Type I error necessary to attain a given power is smaller when using T^1 than when using T^2. This is so because $(\log\alpha^1/\log\alpha^2) > 1$ is equivalent to $\log\alpha^1 < \log\alpha^2$ since $\log\alpha^2$ is negative. Equivalently, the test T^1 is preferred to the test T^2 if T^1 requires fewer observations necessary to attain a given power than T^2 does. It is the latter interpretation that appeared in Table 23.1, where various methods are summarized.

c) Determination of Approximate Slopes

From a practical point of view it remains to explain how the asymptotic exact slope $c(\theta)$ and the approximate slope $c^a(\theta)$ are determined. A result, which is of simple use, is given next for the case of approximate slopes.

Property 23.14: *Consider a sequence of test statistics $\{T_n\}$ satisfying the following hypotheses:*

A1: For every value $\theta \in \Theta_0$, T_n converges in distribution to a random variable whose probability distribution is independent of θ, $\theta \in \Theta_0$, and is continuous.

A2: The cumulative distribution function F of this limiting distribution satisfies $\log(1-F(x)) \# -ax^2/2$ *as x increases to infinity, where a is a strictly positive scalar.*

A3: For every value $\theta \in \Theta_1$, $T_n/\sqrt{n}$ converges to a strictly positive scalar $b(\theta)$ in P_θ-probability.

Then, under these assumptions, we have

$$c^a(\theta) = ab(\theta)^2.$$

SKETCH OF THE PROOF: We have $\Lambda(t) = 1 - F(t)$. Hence

$$\begin{aligned} K_n^a(T_n) &= -\frac{2}{n}\log\Lambda(T_n) \\ &= -\frac{2}{n}\log(1-F(T_n)) \\ &\# -\frac{2}{n}\log(1-F(b(\theta)\sqrt{n})), \quad \forall\theta\in\Theta_1 \\ &\# -\frac{2}{n}\left(-a\frac{(b(\theta)\sqrt{n})^2}{2}\right) \\ &= ab(\theta)^2. \end{aligned}$$

□

A sequence of tests T_n satisfying assumptions A1, A2, and A3 is said to be a *regular sequence*. Thus, if we consider two regular sequences of tests, then Bahadur's ARE is

$$e^B(T^1, T^2, \theta) = \frac{a_1 b_1^2(\theta)}{a_2 b_2^2(\theta)}.$$

In many econometric problems the test statistics are regular with asymptotic distributions that are either normal or chi-squares. Then the preceding formula for Bahadur's ARE simplifies as the next property shows.

Property 23.15:
(i) Let T_n be a positive test statistic such that $(T_n)^2$ converges in distribution to a $\chi^2(r)$ under the null hypothesis. Then the coefficient a is equal to one.

(ii) Suppose that T_n^1 and T_n^2 are regular sequences with $a = 1$. Then Bahadur's ARE is

$$e^B(T^1, T^2, \theta) = \text{plim}_n \left(\frac{(T_n^1)^2/n}{(T_n^2)^2/n} \right) = \text{plim}_n \frac{(T_n^1)^2}{(T_n^2)^2}.$$

PROOF:

(i) Recall that a random variable T^2 follows a $\chi^2(r)$ distribution if and only if $T^2/2$ follows a $\gamma(r/2)$ distribution. Let $r/2 = \rho$. We obtain

$$\begin{aligned} F(x) &= P(T < x) = P\left(\frac{T^2}{2} < \frac{x^2}{2} \right) \\ &= \frac{1}{\Gamma(\rho)} \int_0^{x^2/2} \exp(-t) t^{\rho - 1} dt. \end{aligned}$$

Now, for every strictly positive ϵ and any t sufficiently large, if $\rho \geq 1$ (the proof is similar if $\rho < 1$), then

$$\exp(-t) \leq \exp(-t) t^{\rho - 1} \leq \exp(-(1 - \epsilon)t).$$

It follows that

$$\frac{1}{\Gamma(\rho)} \int_{x^2/2}^{\infty} \exp(-t) \leq 1 - F(x) \leq \frac{1}{\Gamma(\rho)} \int_{x^2/2}^{\infty} \exp(-(1 - \epsilon)t) dt,$$

i.e.

$$\frac{1}{\Gamma(\rho)} \exp\left(-\frac{x^2}{2} \right) \leq 1 - F(x) \leq \frac{1}{(1 - \epsilon)\Gamma(\rho)} \exp\left(-\frac{x^2}{2}(1 - \epsilon) \right),$$

i.e.

$$-\frac{x^2}{2} - \log \Gamma(\rho) \leq \log(1 - F(x)) \leq -\log(1 - \epsilon) - \log \Gamma(\rho) - (1 - \epsilon)\frac{x^2}{2}.$$

From these inequalitites it follows that

$$\lim_{x \to \infty} \frac{\log(1 - F(x))}{-\frac{x^2}{2}} = 1,$$

which implies the desired result.

(ii) The second part immediately follows from (i) and Property 23.14.

□

d) Examples

Example 23.10: Consider the simple linear regression model

$$Y_i = X_i\theta + u_i, \quad i = 1, \ldots, n,$$

where $E(u_i \mid X_i) = 0$ and $V(u_i \mid X_i) = 1$. We are interested in testing the hypothesis H_0: $\theta = 0$.

(i) There exist many test statistics satisfying assumptions A1–A3. For instance, consider an instrumental variable Z, i.e., a variable independent of u. Consider the empirical analog of $\text{cov}(u, Z)$ under the null hypothesis, i.e.

$$\xi_n = \frac{1}{n}\sum_{i=1}^{n} Z_i\hat{u}_i^0 = \frac{1}{n}\sum_{i=1}^{n} Z_iY_i,$$

for, under H_0, $\hat{u}_i^0 = Y_i = u_i$.

Now, under the null hypothesis and assuming that the observations are iid, we have

$$\sqrt{n}\,\xi_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_iu_i \xrightarrow{d} N(E(Zu), V(Zu)),$$

or equivalently

$$\sqrt{n}\,\xi_n \xrightarrow{d} N(0, E(Z^2))$$

because $V(Zu) = E(Zu)^2 = E(Z^2u^2) = E(Z^2E(u^2 \mid Z)) = EZ^2$.

Therefore the test statistic

$$T_n^2(Z) = \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_iY_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}$$

follows a $\chi^2(1)$ distribution under the null hypothesis.

Note that there are at least as many regular test statistics as possible instrumental variables Z. Moreover, note that the exact probability distribution of the test statistic $T_n(Z)$ is unknown since no specific hypothesis is made on the probability distribution of the error terms. As a result, only the approximate method can be used.

(ii) The conditions under which Properties 23.14 and 23.15 hold are satisfied in this example for T_n, which is the positive square root of the above right-hand side. In particular, under the alternative hypothesis,

we have

$$\begin{aligned}\lim_n \frac{T_n^2(Z)}{n} &= \frac{\left(\text{plim}_n \frac{1}{n}\sum_{i=1}^n Z_i Y_i\right)^2}{\text{plim}_n \frac{1}{n}\sum_{i=1}^n Z_i^2} \\ &= \frac{\left(\text{plim}_n \frac{1}{n}\sum_{i=1}^n Z_i X_i \theta + \text{plim}_n \frac{1}{n}\sum_{i=1}^n Z_i u_i\right)^2}{\text{plim}_n \frac{1}{n}\sum_{i=1}^n Z_i^2} \\ &= \frac{(E(ZX))^2\theta^2}{E(Z^2)} \equiv b_Z(\theta)^2.\end{aligned}$$

(iii) Among the preceding tests, the asymptotically optimal test is obtained when $(E^2(ZX)/E(Z^2))\theta^2$ is maximal. The latter quantity is equal to

$$\frac{(E(ZX))^2}{E(Z^2)E(X^2)} E(X^2)\theta^2.$$

Hence, from Cauchy–Schwarz inequality this quantity is maximal if the intrumental variable Z is proportional to X.

Example 23.11:

(i) Consider again the preceding example with the optimal choice X of instrumental variable. Suppose now that the conditional variance of the error term is equal to σ^2. A statistic for testing the null hypothesis H_0: $\theta = 0$ is

$$T_n^2 = \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i Y_i\right)^2}{\tilde{\sigma}_n^2 \frac{1}{n}\sum_{i=1}^n X_i^2},$$

where $\tilde{\sigma}_n^2$ is a consistent estimator of σ^2 under the null hypothesis. Under H_0, this test statistic is asymptotically $\chi^2(1)$ distributed.

Under the alternative hypothesis, we have

$$\text{plim}_n \frac{T_n^2}{n} = \frac{\theta^2 EX^2}{\text{plim}_n \tilde{\sigma}_n^2},$$

where $\text{plim}_n \tilde{\sigma}_n^2$ denotes the probability limit of the variance estimator under the alternative hypothesis. Thus, according to Bahadur's definition of efficiency, one should choose a variance estimator whose limit is as small as possible.

(ii) Next, consider the usual estimator

$$\hat{\sigma}_n^{02} = (1/n)\sum_{i=1}^n Y_i^2$$

of σ^2 under the null hypothesis H_0, which is used in the Lagrange multiplier test, and the estimator

$$\hat{\sigma}_n^{02} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i\left(\sum_{i=1}^{n}X_i^2\right)^{-1}\sum_{i=1}^{n}X_iY_i\right)^2$$

of σ^2 under the maintained hypothesis, which is used in the Wald test.

Under the alternative hypothesis, we have

$$\begin{aligned}\text{plim}_n\hat{\sigma}_n^{02} &= \text{plim}_n\frac{1}{n}\sum_{i=1}^{n}(X_i\theta + u_i)^2 \\ &= E(X\theta + u)^2 \\ &= E(X^2)\theta^2 + Eu^2 + 2\theta E(Xu) \\ &= \sigma^2 + E(X^2)\theta^2\end{aligned}$$

and

$$\lim_n \hat{\sigma}_n^2 = \lim_n \frac{1}{n}\sum_{i=1}^{n}u_i^2 = \sigma^2.$$

Because $\text{plim}_n\hat{\sigma}_n^2 < \text{plim}_n\hat{\sigma}_n^{02}$, it follows that the Wald test is asymptotically more powerful in Bahadur's sense than the Lagrange multiplier test. Intuitively, such a result is expected because the Wald test uses the estimation of variance σ^2 under the alternative hypothesis and, hence, takes greater account of the alternative hypothesis than the Lagrange multiplier test. We emphasize, however, that the preceding result is based on the approximate method. In particular, nothing is said about the relative efficiency of these two tests in finite samples.

23.2.4 Pitman's Relative Efficiency

Within Pitman's framework, competing tests are compared under local alternatives using the dual approach. Thus a parameter value θ_n is associated to each sample size n. Thus the "true value" θ_n is no longer fixed within such an approach. As a result, it is necessary to develop an asymptotic theory based on doubly indexed sequences. This is the purpose of the next subsection, where a generalization of the standard Central Limit Theorem is provided. Then this generalization is used to derive the asymptotic properties of estimators under sequences of local alternatives.

a) A Generalized Central Limit Theorem

To derive the asymptotic properties of interest, we begin with a generalization of the standard Central Limit Theorem to the case of doubly indexed sequences. Let $Y_{i,n}$, $n \in \mathbb{N}$, $i \in \{1, \ldots, n\}$, denote such a sequence. For every n, it is assumed that the variables $Y_{i,n}$ are independently and identically distributed with mean m_n and variance σ_n^2. Then the behavior of $\bar{Y}_n = (1/n)\sum_{i=1}^n Y_{i,n}$ can be studied as in the standard Central Limit Theorem B.8. The characteristic function of $\sqrt{n}(\bar{Y}_n - m_n)$ is

$$\psi_n(t) = E(\exp it\sqrt{n}(\bar{Y}_n - m_n)) = \left(\phi_n\left(\frac{t}{\sqrt{n}}\right)\right)^n,$$

where ϕ_n denotes the characteristic function of $Y_{i,n} - m_n$. Then, if the variance σ_n^2 converges to a limit σ^2, we have

$$\psi_n(t) \# \left(1 - \frac{t^2\sigma_n^2}{2n}\right)^n \to \exp\left(-\frac{t^2\sigma^2}{2}\right),$$

in general. Hence we obtain the following convergence in distribution

$$\sqrt{n}(\bar{Y}_n - m_n) \xrightarrow{d} N(0, \sigma^2).$$

Moreover, if it is assumed that the mean m_n is asymptotically equivalent to $m + \mu/\sqrt{n}$, then we obtain

$$\sqrt{n}(\bar{Y}_n - m) - \mu \xrightarrow{d} N(0, \sigma^2),$$

i.e.

$$\sqrt{n}(\bar{Y}_n - m) \xrightarrow{d} N(\mu, \sigma^2).$$

To summarize, when the parameters of the probability distributions of $Y_{i,n}$ are no longer fixed, we can still obtain a convergence to a normal distribution although the limiting distribution is no longer centered. Note, however, that this results from the chosen rate of convergence in $1/\sqrt{n}$ of the mean m_n.

b) Pitman's Example

Pitman essentially developed his theory for a special case. Specifically, the model is parametric. The null hypothesis is H_0: $\theta = \theta_0$ while the alternative hypothesis is H_1: $\theta > \theta_0$. Suppose that two available test statistics T_n^j, $j = 1, 2$, satisfy

A1: $E_\theta T_n^j = m_n^j(\theta)$ and $V_\theta(T_n^j) = \sigma_n^{j2}(\theta)$, $\forall \theta \in \Theta = [\theta_0, \infty[$,

A2: $\dfrac{dm_n^j(\theta_0)}{d\theta} > 0$,

A3: $\lim_{n\to\infty} \dfrac{1}{\sqrt{n}} \dfrac{dm_n^j(\theta_0)}{d\theta} / \sigma_n^j(\theta_0) = c_j > 0$,

A4: For every sequence of local alternatives H_{1n}: $\theta = \theta_0 + (\mu/\sqrt{n})$, $\mu > 0$, we have:

$$(i) \quad \lim_{n\to\infty} \frac{dm_n^j\left(\theta_0 + \frac{\mu}{\sqrt{n}}\right)}{d\theta} / \frac{dm_n^j(\theta_0)}{d\theta} = 1,$$

$$(ii) \quad \lim_{n\to\infty} \sigma_n^j\left(\theta_0 + \frac{\mu}{\sqrt{n}}\right) / \sigma_n^j(\theta_0) = 1,$$

$$(iii) \quad \frac{T_n^j - m_n^j\left(\theta_0 + \frac{\mu}{\sqrt{n}}\right)}{\sigma_n^j\left(\theta_0 + \frac{\mu}{\sqrt{n}}\right)} \xrightarrow{d} N(0,1) \text{ under the sequence } H_{1n}.$$

Consider now the approximate method. To each of the above test statistic we can associate a critical region

$$W^{aj}(n,\alpha) = \{T_n^j \geq m_n^j(\theta_0) + \Phi^{-1}(1-\alpha)\sigma_n^j(\theta_0)\} \tag{23.15}$$

whose risk of Type I error is asymptotically equal to α.

Property 23.16: *Consider two test statistics T_n^1 and T_n^2 satisfying assumptions A1–A4 for every $\mu \geq 0$. Let $W^{aj}(n,\alpha)$, $j = 1,2$ be the corresponding critical regions. Then we have*

$$\lim_{n\to\infty} \frac{N_2\left(\alpha,\gamma,\theta_0 + \frac{\mu}{\sqrt{n}}\right)}{N_1\left(\alpha,\gamma,\theta_0 + \frac{\mu}{\sqrt{n}}\right)} = \left(\frac{c_1}{c_2}\right)^2 = \lim_{n\to\infty} \left(\frac{\frac{dm_n^1(\theta_0)}{d\theta} / \sigma_n^1(\theta_0)}{\frac{dm_n^2(\theta_0)}{d\theta} / \sigma_n^2(\theta_0)} \right)^2 .$$

SKETCH OF THE PROOF: The basic idea of the proof is that the minimal number of observations N is asymptotically proportional to the number of observations n. Then the desired result follows from a series expansion of the power function, when the sequence of alternatives θ_n converges to θ_0, and when the minimal number of observations N increases to infinity proportionally to n.

More precisely, we have

$$\begin{aligned}
&\gamma^j\left(N,\alpha,\theta_0+\frac{\mu}{\sqrt{n}}\right)\\
&\quad= P_{\theta_0+\mu/\sqrt{n}}(T_N^j > m_N^j(\theta_0)+\Phi^{-1}(1-\alpha)\sigma_N^j(\theta_0))\\
&\quad= P_{\theta_0+\mu/\sqrt{n}}\left(\frac{T_N^j - m_N^j\left(\theta_0+\frac{\mu}{\sqrt{n}}\right)}{\sigma_N^j\left(\theta_0+\frac{\mu}{\sqrt{n}}\right)}\right.\\
&\qquad\left.> \frac{m_N^j(\theta_0) - m_N^j\left(\theta_0+\mu/\sqrt{n}\right)+\Phi^{-1}(1-\alpha)\sigma_N^j(\theta_0)}{\sigma_N^j\left(\theta_0+\mu/\sqrt{n}\right)}\right)
\end{aligned}$$

Now, take $N \# \lambda n$. Then we obtain

$$\begin{aligned}
&\gamma^j\left(N,\alpha,\theta_0+\frac{\mu}{\sqrt{n}}\right)\\
&\quad\# \quad P_{\theta_0+\mu\sqrt{\lambda}/\sqrt{N}}\left(\frac{T_N^j - m_N^j\left(\theta_0+\mu\sqrt{\lambda}/\sqrt{N}\right)}{\sigma_N^j\left(\theta_0+\mu\sqrt{\lambda}/\sqrt{N}\right)}\right.\\
&\qquad\left.> -\frac{\mu\sqrt{\lambda}}{\sqrt{N}}\frac{dm_N^j(\theta_0)}{d\theta}\frac{1}{\sigma_N^j(\theta_0)}+\Phi^{-1}(1-\alpha)\right)\\
&\quad\# \quad 1-\Phi\left(-\frac{\mu\sqrt{\lambda}}{\sqrt{N}}\frac{dm_N^j(\theta_0)}{d\theta}\frac{1}{\sigma_N^j(\theta_0)}+\Phi^{-1}(1-\alpha)\right)\\
&\quad\# \quad \Phi\left(\mu\sqrt{\lambda}c_j-\Phi^{-1}(1-\alpha)\right).
\end{aligned}$$

Thus, asymptotically, we have

$$N_j\left(\alpha,\gamma,\theta_0+\frac{\mu}{\sqrt{n}}\right) \# \lambda_j n,$$

where λ_j satisfies

$$\Phi\left(\mu\sqrt{\lambda_j}c_j-\Phi^{-1}(1-\alpha)\right)=\gamma,$$

i.e.

$$\sqrt{\lambda_j}=\frac{\Phi^{-1}(1-\alpha)+\Phi^{-1}(\gamma)}{\mu c_j},$$

i.e.

$$\lambda_j=\left(\frac{\Phi^{-1}(1-\alpha)+\Phi^{-1}(\gamma)}{\mu c_j}\right)^2.$$

Hence we have

$$N_j\left(\alpha,\gamma,\theta_0+\frac{\mu}{\sqrt{n}}\right) \# \left(\frac{\Phi^{-1}(1-\alpha)+\Phi^{-1}(\gamma)}{\mu}\right)^2 \frac{n}{c_j^2}.$$

Therefore

$$\lim_{n\to\infty}\frac{N_2\left(\alpha,\gamma,\theta_0+\mu/\sqrt{n}\right)}{N_1\left(\alpha,\gamma,\theta_0+\mu/\sqrt{n}\right)}=\left(\frac{c_1}{c_2}\right)^2.$$

□

By definition, the limit obtained in Property 23.16 is the *asymptotic relative efficiency in Pitman's sense.* That is

$$e^p(T_1,T_2)=\lim_{n\to\infty}\frac{N_2\left(\alpha,\gamma,\theta_0+\frac{\mu}{\sqrt{n}}\right)}{N_1\left(\alpha,\gamma,\theta_0+\frac{\mu}{\sqrt{n}}\right)} \tag{23.16}$$

Note that in the special case considered this quantity is independent of α, γ, and the value μ defining the sequence of local alternatives.

Remark 23.2: The computation of the limits of the power functions in the proof of Property 23.16 gives

$$\lim_{n\to\infty}\gamma^j\left(n,\alpha,\theta_0+\frac{\mu}{\sqrt{n}}\right)=\Phi(\mu c_j-\Phi^{-1}(1-\alpha)),$$

where $\mu>0$. A direct comparison of these power functions leads to the choice of the test that corresponds to the largest c_j, i.e., the test that is locally more powerful. In general, direct computation is simpler than the use of the minimal number of observations $N_j(\alpha,\gamma,\theta_0+\mu/\sqrt{n})$.

23.2.5 Direct Approach Based on Local Alternatives

Alternatively, we can compare directly the competing testing procedures in the neighborhood of the null hypothesis by means of their power functions. The approach, which is usually followed in econometrics, consists in determining first the asymptotic distribution of the estimator of the parameter under the given sequence of local alternatives. Then one derives the asymptotic distribution of the corresponding test statistic, and, hence, the corresponding form of the power function.

In this subsection we shall illustrate such an approach when the competing tests are based on M-estimators. We consider observations (Y_{in},X_{in}), $i=1,\ldots,n$, $n\in\mathbb{N}$, that are doubly indexed. For every value

of n, the observations are assumed to be independently and identically distributed P_{θ_n}, where $\theta_n = \theta_0 + \mu/\sqrt{n}$, and θ_0 and μ belong to $\mathbb{R}^p$. We consider the M-estimator defined as the solution $\tilde{\theta}_n$ to

$$\min_{\theta \in \Theta} \sum_{i=1}^{n} \psi(Y_{in}, X_{in}; \theta).$$

a) Consistency

As usual, we begin with the consistency of $\tilde{\theta}_n$ by studying the corresponding limit optimization problem. Under suitable regularity conditions, we have

$$\frac{1}{n} \sum_{i=1}^{n} \psi(Y_{in}, X_{in}; \theta) \# E_{\theta_n} \psi(Y, X; \theta) \# E_{\theta_0} \psi(Y, X; \theta).$$

If the limit optimization problem $\min_{\theta \in \Theta} E_{\theta_0} \psi(Y, X; \theta)$ admits θ_0 as a unique solution, then the M-estimator $\tilde{\theta}_n$ exists and is consistent for θ_0. In other words, a local theory in the neighborhood of θ_0 does not invalidate the consistency property of $\tilde{\theta}_n$.

b) Asymptotic Distribution

To derive the asymptotic distribution of $\tilde{\theta}_n$, it is necessary to expand the corresponding first-order conditions. From

$$\sum_{i=1}^{n} \frac{\partial \psi(Y_{in}, X_{in}; \tilde{\theta}_n)}{\partial \theta} = 0,$$

we obtain

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \psi(Y_{in}, X_{in}; \theta_n)}{\partial \theta} + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \psi(Y_{in}, X_{in}; \theta_n)}{\partial \theta \partial \theta'} \sqrt{n}(\tilde{\theta}_n - \theta_n) \# 0,$$

i.e.

$$\sqrt{n}(\tilde{\theta}_n - \theta_n) \# \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \psi(Y_{in}, X_{in}; \theta_n)}{\partial \theta \partial \theta'} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \psi(Y_{in}, X_{in}; \theta_n)}{\partial \theta}$$
$$\# \left[E_{\theta_0} \left(\frac{\partial^2 \psi(Y, X; \theta_0)}{\partial \theta \partial \theta'} \right) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \psi(Y_{in}, X_{in}; \theta_n)}{\partial \theta}.$$

When the common probability distribution is P_{θ_n}, the random vectors $\partial\psi(Y_{in}, X_{in}; \theta_n)/\partial\theta$ are independently and identically distributed with zero mean and variance covariance matrix Σ_n converging to

$$E_{\theta_0}\left(\partial\psi(Y, X; \theta_0)/\partial\theta \times \partial\psi(Y, X, \theta_0)/\partial\theta'\right).$$

Hence, from a multivariate version of the generalized Central Limit Theorem of Section 13.2.4-a) it follows that

$$\sqrt{n}(\tilde{\theta}_n - \theta_n) \xrightarrow{d} N(0, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}),$$

where

$$\begin{aligned} \mathcal{J} &= E_{\theta_0}\left(-\frac{\partial^2\psi(Y, X; \theta_0)}{\partial\theta\partial\theta'}\right), \\ \mathcal{I} &= E_{\theta_0}\left(\frac{\partial\psi(Y, X; \theta_0)}{\partial\theta}\frac{\partial\psi(Y, X; \theta_0)}{\partial\theta'}\right). \end{aligned} \tag{23.17}$$

Replacing θ_n by $\theta_0 + \mu/\sqrt{n}$, we obtain the following property.

Property 23.17: *Under regularity conditions generalizing those of Chapter 8 and under the sequence of local alternatives $\theta_n = \theta_0 + \mu/\sqrt{n}$, the M-estimator $\tilde{\theta}_n$ satisfies*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} N(\mu, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}),$$

where $\mathcal{J}$ and $\mathcal{I}$ are given by equation (23.17).

The asymptotic distribution of the M-estimator $\tilde{\theta}_n$ has therefore a nonzero mean. As a result, quadratic forms based on $\tilde{\theta}_n$ will also exhibit a noncentrality parmater.

Corollary 23.1: *For the M-estimator $\tilde{\theta}_n$ we have*

$$n(\tilde{\theta}_n - \theta_0)'\mathcal{J}\mathcal{I}^{-1}\mathcal{J}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \chi^2(p, \delta),$$

where the noncentrality parameter of the chi-square distribution is

$$\delta = \mu'\mathcal{J}\mathcal{I}^{-1}\mathcal{J}\mu.$$

Proof: This is an immediate consequence of Property B.49. □

Corollary 23.1 gives directly the asymptotic distributions of commonly used test statistics under sequences of local alternatives. For instance, suppose that we want to test the null hypothesis H_0: $\theta = \theta_0$. Then a natural test statistic is

$$\xi_n = n(\tilde{\theta}_n - \theta_0)' \hat{\mathcal{J}} \hat{\mathcal{I}}^{-1} \hat{\mathcal{J}} (\tilde{\theta}_n - \theta_0), \tag{23.18}$$

where $\hat{\mathcal{J}}$ and $\hat{\mathcal{I}}$ are consistent estimators of $\mathcal{J}$ and $\mathcal{I}$, respectively. Then, it follows from Corollary 23.1 that the probability distribution of ξ_n can be approximated by a $\chi^2(p, \delta)$ distribution under the sequence of local alternatives H_{1n}: $\theta = \theta_0 + \mu/\sqrt{n}$.

Hence an approximate critical region at level α based on such a test statistic is

$$W(n, \alpha) = \{\xi_n > \chi^2_{1-\alpha}(p)\}, \tag{23.19}$$

where $\chi^2_{1-\alpha}(p)$ denotes the $(1-\alpha)$-th quantile of the chi-square distribution with p degrees of freedom.

Moreover, under this sequence of local alternatives, the asymptotic power function is

$$\lim_n \gamma\left(n, \alpha, \theta_0 + \frac{\mu}{\sqrt{n}}\right) = Q_{p,\delta}([\chi^2_{1-\alpha}(p), +\infty)), \tag{23.20}$$

where $Q_{p,\delta}$ is the distribution function of the $\chi^2(p, \delta)$ distribution.

This power function allows us to compare easily competing testing procedures. Specifically, noting that the cumulative distribution function of a $\chi^2(p, \delta)$ distribution is a decreasing function of the noncentrality parameter δ, we obtain at once the following corollary.

Corollary 23.2: *Consider two tests of the hypothesis H_0: $\theta = \theta_0$ based on the M-estimators $\tilde{\theta}^1_n$ and $\tilde{\theta}^2_n$. Let ξ^1_n and ξ^2_n denote the corresponding test statistics. Suppose that*

$$\xi^1_n \to \chi^2(p_1, \delta_1) \quad \text{and} \quad \xi^2_n \to \chi^2(p_1, \delta_2)$$

under the sequence of local alternatives $\theta_n = \theta_0 + \mu/\sqrt{n}$. If $\delta_1 > \delta_2$, then the first test is preferred to the second test under this sequence of local alternatives.

In some cases the dominance of a test over another test is uniform in the auxiliary parameter μ defining the sequence of local alternatives. For instance, this arises when

$$\delta_1 = \mu' \mathcal{J}_1 \mathcal{I}_1^{-1} \mathcal{J}_1 \mu \geq \delta_2 = \mu' \mathcal{J}_2 \mathcal{I}_2^{-1} \mathcal{J}_2 \mu, \quad \forall \mu,$$

i.e.

$$\mathcal{J}_1\mathcal{I}_1^{-1}\mathcal{J}_1 \succeq \mathcal{J}_2\mathcal{I}_2^{-1}\mathcal{J}_2,$$

i.e.

$$\mathcal{J}_1^{-1}\mathcal{I}_1\mathcal{J}_1^{-1} \preceq \mathcal{J}_2^{-1}\mathcal{I}_2\mathcal{J}_2^{-1}.$$

Because $\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}$ is the asymptotic variance covariance matrix of the M-estimator $\tilde{\theta}_n$ under the null hypothesis, it follows that the first test is uniformly preferred to the second test under sequences of local alternatives if and only if the estimator on which it is based is asymptotically more efficient than the estimator on which the second test is based. Therefore this result relates the optimality of testing procedures to the efficiency of estimators.

The preceding results were established for the special case where the null hypothesis is of the form $\theta = \theta_0$. It is, however, easy to generalize these results to null hypotheses defined by equality constraints in implicit, explicit, or mixed forms.

Similarly, it is useful to note that the asymptotic expansions of the various test statistics that were obtained in Chapters 17 and 18 around the true value θ_0, when θ_0 is assumed to belong to the null hypothesis remain valid when they are derived around $\theta_n = \theta_0 + (\mu/\sqrt{n})$. In particular, it follows that the asymptotic equivalences among Wald type, score, Lagrange multiplier, and log-likelihood ratio test statistics obtained in these chapters under the null hypothesis remain valid under sequences of local alternatives converging to θ_0.

23.3 Semiparametric Efficiency

In the preceding sections the asymptotic optimality properties of the various inference procedures were studied within a parametric framework. Given that semiparametric models are frequently encountered in econometrics, it is useful to extend these properties to such models. In this section we present an extension of the concept of first-order efficiency, which recently appears in the econometric literature.

23.3.1 Information Contained in Identifying Conditions

Let (X_i, Y_i), $i = 1, \ldots, n$, be independently and identically distributed observations with common probability distribution P. Let $\theta \in \Theta \subset \mathbb{R}^p$

be the parameter vector of interest. We consider a semiparametric model defined by constraints of the form

$$E_P(g(Y, X; \theta) \mid X) = 0, \tag{23.21}$$

where g is a function with values in $\mathbb{R}^K$ assumed to be differentiable with respect to θ.

More precisely, the family of probability distributions associated with a given value θ of the parameter is

$$\mathcal{P}_\theta = \{P : E_P(g(Y, X; \theta) \mid X) = 0\}.$$

In general, this family contains an infinity of probability distributions. This constitutes the semiparametric aspect of the problem since a value of the parameter is insufficient to characterize a probability distribution. However, the knowledge that the parameter θ is defined through the above constraints associated with g actually contains information about the parameter θ. Thus our goal is to construct a measure of such information, which is analogous to the Fisher information measure introduced in the parametric case. A priori, such measure should depend upon the semiparametric model and the probability distribution retained in this model.

A natural method, which makes a link between parametric and semiparametric situations, consists in considering all the parametric models that are compatible with the specified semiparametric model. Such a parametric model is

$$\tilde{\mathcal{P}}_{\Theta,B} = \{f(y \mid x; \theta, \beta) f_0(x), (\theta, \beta) \in \Theta \times B\},$$

where the probability density $f(y \mid x, \theta, \beta) f_0(x)$ belongs to the family $\mathcal{P}_\theta$ for every given value of θ. As a matter of fact, this model is parametric only for any given probability density f_0.

Consider now a probability distribution P associated with a value θ of the parameter so that $P \in \mathcal{P}_\theta$. Consider also a parametric model such that $P \in \tilde{\mathcal{P}}_{\theta,B} \subset \mathcal{P}_\theta$. To the probability distribution P there corresponds a value β of the additional parameter. Information about the parameter vector (θ, β) contained in the parametric model, when the true probability distribution is P, is summarized in the Fisher information matrix

$$\mathcal{I}_P(\theta, \beta) = E_P \begin{pmatrix} \dfrac{-\partial^2 \log f(Y \mid X; \theta, \beta)}{\partial\theta\partial\theta'} & \dfrac{-\partial^2 \log f(Y \mid X; \theta, \beta)}{\partial\theta\partial\beta'} \\ \dfrac{-\partial^2 \log f(Y \mid X; \theta, \beta)}{\partial\beta\partial\theta'} & \dfrac{-\partial^2 \log f(Y \mid X; \theta, \beta)}{\partial\beta\partial\beta'} \end{pmatrix}$$

$$= \begin{pmatrix} \mathcal{I}_{\theta\theta} & \mathcal{I}_{\theta\beta} \\ \mathcal{I}_{\beta\theta} & \mathcal{I}_{\beta\beta} \end{pmatrix}.$$

The FDCR bound associated with the parameter subvector θ is the top-left block of the inverse of the information matrix, which we denote by $\mathcal{I}^{\theta\theta}$. Thus we see that a natural measure of information about θ contained in the parametric submodel is

$$(\mathcal{I}^{\theta\theta})^{-1} = \mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\theta},$$

when the true probability distribution is P.

This leads to the following definition.

Definition 23.9: *The information about θ contained in the semiparametric model (23.21), when the true probability distribution is P is the minimum of the information contained in any compatible parametric submodels containing P. That is, if $\mathcal{I}_{g,P}(\theta)$ denotes this information, then*

$$\mathcal{I}_{g,P}(\theta) = \min_{\tilde{\mathcal{P}}_{\Theta,B}} \left(\mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\theta}\right),$$

where the minimum is taken relative to the usual preorder on symmetric matrices.

We now determine such a bound. This is done in two steps.

STEP 1: Consider a compatible parametric submodel containing P. We have $E_P(g(Y,X;\theta) \mid X) = 0$ for any probability distribution belonging to the parametric model. That is

$$\int_{\mathcal{Y}} g(y,x;\theta) f(y \mid x;\theta,\beta) d\mu(y) = 0, \quad \forall \theta,\beta,x.$$

Differentiating this equality with respect to the parameters β and θ, we obtain

$$\mathrm{Cov}_P\left(g(Y,X;\theta), \frac{\partial \log f(Y \mid X;\theta,\beta)}{\partial \beta} \mid X\right) = 0,$$

$$E_P\left(\frac{\partial g(Y,X;\theta)}{\partial \theta'} \mid X\right) + \mathrm{Cov}_P\left(g(Y,X;\theta), \frac{\partial \log f(Y \mid X;\theta,\beta)}{\partial \theta} \mid X\right) = 0.$$

But, from Property B.20 we have

$$V_P\left(\begin{pmatrix} \frac{\partial \log f}{\partial \theta} \\ \frac{\partial \log f}{\partial \beta} \end{pmatrix} \mid X\right) \succeq$$

$$\mathrm{Cov}_P\left(\begin{pmatrix} \frac{\partial \log f}{\partial \theta} \\ \frac{\partial \log f}{\partial \beta} \end{pmatrix}, g \mid X\right) V_P(g \mid X)^{-1} \mathrm{Cov}_P\left(g, \begin{pmatrix} \frac{\partial \log f}{\partial \theta} \\ \frac{\partial \log f}{\partial \beta} \end{pmatrix} \mid X\right).$$

It follows that

$$V_P\left(\begin{pmatrix} \frac{\partial \log f}{\partial \theta} \\ \frac{\partial \log f}{\partial \beta} \end{pmatrix} \mid X\right) \succeq$$
$$\begin{pmatrix} E_P\left(\frac{\partial g'}{\partial \theta} \mid X\right) \\ 0 \end{pmatrix} V_P(g \mid X)^{-1} \left(E_P\left(\frac{\partial g}{\partial \theta'} \mid X\right), 0\right).$$

Integrating with respect to the marginal probability distribution of X, we obtain the inequality

$$\begin{pmatrix} \mathcal{I}_{\theta\theta} & \mathcal{I}_{\theta\beta} \\ \mathcal{I}_{\beta\theta} & \mathcal{I}_{\beta\beta} \end{pmatrix} \succeq$$
$$\begin{pmatrix} E_P\left(E_P\left(\frac{\partial g'}{\partial \theta} \mid X\right) V_P(g \mid X)^{-1} E_P\left(\frac{\partial g}{\partial \theta'} \mid X\right)\right) & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence

$$\mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\theta} \succeq E_P\left(E_P\left(\frac{\partial g'}{\partial \theta} \mid X\right) V_P(g \mid X)^{-1} E_P\left(\frac{\partial g}{\partial \theta'} \mid X\right)\right). \tag{23.22}$$

Therefore we have proved that

$$\mathcal{I}_{g,P}(\theta) \succeq E_P\left(E_P\left(\frac{\partial g'}{\partial \theta} \mid X\right) V_P(g \mid X)^{-1} E_P\left(\frac{\partial g}{\partial \theta'} \mid X\right)\right).$$

STEP 2: It remains to determine when the above lower bound is attained, for instance by a parametric submodel parameterized by θ only. The lower bound is attained when

$$V_P\left(\frac{\partial \log f}{\partial \theta} \mid X\right) =$$
$$\mathrm{Cov}_P\left(\frac{\partial \log f}{\partial \theta}, g \mid X\right) V_P(g \mid X)^{-1} \mathrm{Cov}_P\left(g, \frac{\partial \log f}{\partial \theta} \mid X\right),$$

i.e., when

$$\frac{\partial \log f}{\partial \theta} = E_P\left(-\frac{\partial g'}{\partial \theta} \mid X\right) V_P(g \mid X)^{-1} g.$$

Hence we have the following property.

Property 23.18: *If it is possible to construct a compatible parametric submodel with densities satisfying*

$$\frac{\partial \log f}{\partial \theta} = E_P\left(-\frac{\partial g'}{\partial \theta} \mid X\right) V_P(g \mid X)^{-1} g,$$

then the information contained in the semiparametric model (23.21) is

$$\mathcal{I}_{g,P}(\theta) = E_P\left(E_P\left(\frac{\partial g'}{\partial \theta} \mid X\right) V_P(g \mid X)^{-1} E_P\left(\frac{\partial g}{\partial \theta'} \mid X\right)\right).$$

Intuitively, the parametric submodel given in Property 23.18 is the parametric model that is the least informative about θ while still being compatible with the constraints (23.21).

Example 23.12: Consider the problem of estimating the parameter vector θ in the regression model

$$E(Y - m(X;\theta) \mid X) = 0,$$

where Y and m are univariate. We have

$$g(X, Y; \theta) = Y - m(X;\theta),$$
$$V_P(g \mid X) = V_P(Y \mid X) \equiv \sigma^2(X),$$
$$E_P\left(-\frac{\partial g}{\partial \theta} \mid X\right) = \frac{\partial m(X;\theta)}{\partial \theta}.$$

The condition of Property 23.18 is

$$\frac{\partial \log f(y \mid x;\theta)}{\partial \theta} = \frac{\partial m(x;\theta)}{\partial \theta} \frac{y - m(x;\theta)}{\sigma^2(x)}.$$

It is clear that the family of normal distributions $N(m(x;\theta), \sigma^2(x))$ satisfies the preceding condition while also being compatible with the semiparametric model. Hence the semiparametric information bound is

$$\mathcal{I}_{g,P} = E_P\left(\frac{1}{\sigma^2(X)} \frac{\partial m(X;\theta)}{\partial \theta} \frac{\partial m(X;\theta)}{\partial \theta'}\right). \tag{23.23}$$

Example 23.13: Equation (23.22) is especially useful for determining the semiparametric bound in a model written in the implicit form

$$h(Y, X, u; \theta) = 0, \text{ with } E(u \mid X) = 0, V(u \mid X) = \Omega(X).$$

Then the function g is obtained by solving the above system. That is, g satisfies

$$h(Y, X, g(Y, X; \theta); \theta) = 0.$$

From the Implicit Function Theorem, we obtain

$$\frac{\partial g(Y, X; \theta)}{\partial \theta'} = -\left(\frac{\partial h(Y, X, g(Y, X; \theta); \theta)}{\partial u'}\right)^{-1} \frac{\partial h(Y, X, g(Y, X; \theta); \theta)}{\partial \theta'}.$$

Hence the semiparametric information bound is

$$\mathcal{I}_{g,P}(\theta) =$$
$$E_P\left[E_P\left(\left(\frac{\partial h(X, Y, g(Y, X; \theta); \theta)}{\partial u'}\right)^{-1} \frac{\partial h(X, Y, g(Y, X; \theta); \theta)}{\partial \theta'} \mid X\right)'\right.$$
$$\left.\Omega(X)^{-1} E_P\left(\left(\frac{\partial h(X, Y, g(Y, X; \theta); \theta)}{\partial u'}\right)^{-1} \frac{\partial h(X, Y, g(Y, X; \theta); \theta)}{\partial \theta'} \mid X\right)\right].$$

It remains, however, to examine whether such a bound can be attained by a compatible parametric model.

Example 23.14: A similar approach can be applied to a semiparametric model that specifies only the form of the conditional median. As usual in this case, one cannot reverse the expectation operator and the differentiation operator because the function g is not differentiable. Nonetheless, in this case the semiparametric information bound is given by Property 23.18, where $E\left((\partial g'/\partial \theta) \mid X\right)$ is replaced by $\partial E(g' \mid X)/\partial \theta$.

The identifying constraint is

$$E\left(\mathbb{I}_{Y - m(X;\theta) < 0} - (1/2)\right) = 0.$$

Thus we have

$$g = E\mathbb{I}_{Y - m(X;\theta) < 0} - (1/2),$$
$$E_P(g(X, Y; \theta) \mid X) = P(Y < m(X, \theta)) - \frac{1}{2},$$
$$\left(\frac{\partial E_P(g(X, Y; \theta) \mid X)}{\partial \theta}\right)_{\theta = \theta_0} = \frac{\partial m(X, \theta_0)}{\partial \theta} f_X(0),$$

where f_X is the density of the probability distribution of $Y - m(X, \theta_0)$ conditional upon X. Moreover

$$V_P(g \mid X) = V_P\left(\mathbb{I}_{Y - m(X,\theta) < 0} \mid X\right) = \frac{1}{4}.$$

The condition given in Property 23.18 on the score of the least favorable parametric model becomes

$$\frac{\partial \log f(y \mid x;\theta)}{\partial \theta} = \frac{\partial m(x,\theta)}{\partial \theta} 4 f_X(0) \left(\mathbb{I}_{Y-m(x;\theta)<0} - \frac{1}{2} \right).$$

It is clear that the Laplace distribution

$$f(y \mid x;\theta) = K_x \exp -2 f_X(0) |y - m(x,\theta)|$$

satisfies this condition. Hence the semiparametric information bound is attained. The latter is given by

$$\mathcal{I}_{g,P}(\theta) = E_P \left(4 f_X(0)^2 \frac{\partial m(X,\theta)}{\partial \theta} \frac{\partial m(X,\theta)}{\partial \theta'} \right). \tag{23.24}$$

23.3.2 Asymptotically Efficient Semiparametric Estimators

Let T_n be a consistent estimator of θ in the semiparametric model defined by the constraints (23.21). In particular, this estimator is also consistent in any compatible parametric submodel $\tilde{\mathcal{P}}_{\Theta,B}$. It is, however, not as efficient as the maximum likelihood estimator of the parametric submodel. Thus, for every $P \in \tilde{\mathcal{P}}_{\Theta,B}$, we have

$$V_{as,P} \sqrt{n}(T_n - \theta) \succeq \left(\mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\beta} \mathcal{I}_{\beta\beta}^{-1} \mathcal{I}_{\beta\theta} \right)^{-1}.$$

Because this inequality holds for every parametric submodel, we have

$$V_{as,P} \sqrt{n}(T_n - \theta) \succeq \max_{\tilde{\mathcal{P}}_{\Theta,B} \ni P} \left(\mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\beta} \mathcal{I}_{\beta\beta}^{-1} \mathcal{I}_{\beta\theta} \right)^{-1},$$

i.e.

$$V_{as,P} \sqrt{n}(T_n - \theta) \succeq \left(\min_{\tilde{\mathcal{P}}_{\Theta,B} \ni P} \left(\mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\beta} \mathcal{I}_{\beta\beta}^{-1} \mathcal{I}_{\beta\theta} \right)^{-1} \right)^{-1} = (\mathcal{I}_{g,P}(\theta))^{-1}$$

Thus we have established the following property.

Property 23.19:
(i) A $\sqrt{n}$-consistent estimator of θ in the semiparametric model (23.21) satisfies

$$V_{as,P} \sqrt{n}(T_n - \theta) \succeq (\mathcal{I}_{g,P}(\theta))^{-1}.$$

(ii) An estimator is asymptotically efficient in the semiparametric model (23.21) if

$$\begin{aligned} V_{as,P}\sqrt{n}(T_n - \theta) &= (\mathcal{I}_{g,P}(\theta))^{-1} \\ &= \left(E_P\left(E_P\left(\frac{\partial g'}{\partial \theta}\,|\,X\right) V_P(g\,|\,X)^{-1} E_P\left(\frac{\partial g}{\partial \theta'}\,|\,X\right)\right)\right)^{-1}. \end{aligned}$$

Example 23.15: In many common cases it is easy to propose efficient estimators. For instance, consider the regression model of Example 23.12. We noted that the least favorable family of probability distributions is the family of normal ditributions with mean $m(X,\theta)$ and variance $\sigma^2(X)$. In addition, one can often construct a consistent estimator $\hat{\sigma}_n^2(X)$ of the conditional variance from the OLS residuals using either a parametric or a nonparametric method. Then a quasi generalized least squares method based on the optimization problem

$$\min_{\theta} \sum_{i=1}^{n} \frac{(Y_i - m(X_i,\theta))^2}{\hat{\sigma}_n^2(X_i)}$$

provides an asymptotically efficient semiparametric estimator. In fact, this estimator is a kind of maximum likelihood estimator based on an estimator of the least favorable parametric family.

A similar approach can be used for estimating efficiently the parameters in the conditional median. In this case the quantity $f_X(0)$ needs to be estimated consistently. Then an efficient estimator is obtained through an estimation method based on weighted least absolute deviations, or equivalently a maximum likelihood method based on the estimated least favorable Laplace distributions.

23.3.3 A Two-Step Method For Obtaining an Asymptotically Efficient Estimator

In the parametric case we know that an asymptotically efficient estimator can be obtained from a $\sqrt{n}$-consistent estimator after one iteration of the Newton–Raphson algorithm. A similar approach can be applied in the semiparametric case.

Specifically, the functions appearing in the information bound are

$$E_P\left(-\frac{\partial g'(Y,X;\theta)}{\partial \theta}\,|\,X\right)$$

and

$$V_P(g(Y, X; \theta) \mid X).$$

In general, these two functions, which depend on the true probability distribution P, do not depend on θ only. Thus suppose that these functions depend on additional parameters α and β so that they can be written as

$$h(X; \theta, \alpha) \quad \text{and} \quad \Omega(X; \theta, \beta),$$

respectively.

Suppose now that θ_n^* is a $\sqrt{n}$–consistent estimator of θ. Suppose also that some consistent estimators α_n^* and β_n^* of α and β are available.

By analogy with the first iteration of the Newton–Raphson algorithm, we can define the modified estimator

$$\begin{aligned}\tilde{\theta}_n &= \theta_n^* + \left(\sum_{i=1}^n h(X_i; \theta_n^*, \alpha_n^*)\Omega^{-1}(X_i; \theta_n^*, \beta_n^*)h'(X_i; \theta_n^*, \alpha_n^*)\right)^{-1} \\ &\quad \times \sum_{i=1}^n h(X_i; \theta_n^*, \alpha_n^*)\Omega^{-1}(X_i; \theta_n^*, \beta_n^*)g(Y_i, X_i; \theta_n^*).\end{aligned}$$

Using the equality $E(g(Y, X; \theta) \mid X) = 0$ and the consistency of the initial estimators, we obtain

$$\begin{aligned}&\sqrt{n}(\tilde{\theta}_n - \theta) \\ &\quad \# \quad \sqrt{n}(\theta_n^* - \theta) + \left(E_P(h(X; \theta, \alpha)\Omega^{-1}(X; \theta, \beta)h'(X; \theta, \alpha))\right)^{-1} \\ &\qquad \times \frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i; \theta, \alpha)\Omega^{-1}(X_i; \theta, \beta) \\ &\qquad\quad \left(g(Y_i, X_i; \theta) + \frac{\partial g(Y_i, X_i; \theta)}{\partial \theta'}(\theta_n^* - \theta)\right) \\ &\quad \# \quad \sqrt{n}(\theta_n^* - \theta) + \left(E(h(X; \theta, \alpha)\Omega^{-1}(X; \theta, \beta)h'(X; \theta, \alpha))\right)^{-1} \\ &\qquad \times \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i; \theta, \alpha)\Omega^{-1}(X_i; \theta, \beta)g(Y_i, X_i; \theta)\right. \\ &\qquad\quad \left. + E\left(h(X; \theta, \alpha)\Omega^{-1}(X; \theta, \beta)\frac{\partial g(Y, X; \theta)}{\partial \theta'}\right)\sqrt{n}(\theta_n^* - \theta)\right].\end{aligned}$$

But

$$E\left(h(X; \theta, \alpha)\Omega^{-1}(X; \theta, \beta)\frac{\partial g(Y, X; \theta)}{\partial \theta'}\right)$$

$$= E\left(h(X;\theta,\alpha)\Omega^{-1}(X;\theta,\beta)E\left(\frac{\partial g(Y,X;\theta)}{\partial\theta'} \mid X\right)\right)$$
$$= -E\left(h(X;\theta,\alpha)\Omega^{-1}(X;\theta,\beta)h'(X;\theta,\alpha)\right).$$

Hence

$$\sqrt{n}(\tilde{\theta}_n - \theta) \quad \# \quad \left(E(h(X;\theta,\alpha)\Omega^{-1}(X;\theta,\beta)h'(X;\theta,\alpha))\right)^{-1} \times \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(X_i;\theta,\alpha)\Omega^{-1}(X_i;\theta,\beta)g(Y_i,X_i;\theta).$$

Under usual regularity conditions, the right-hand side is asymptotically normally distributed with zero mean and variance covariance matrix

$$\begin{aligned}
&(\mathcal{I}_{g,P}(\theta))^{-1}V_P\left(h(X;\theta,\alpha)\Omega^{-1}(X;\theta,\beta)g(Y,X;\theta)\right)(\mathcal{I}_{g,P}(\theta))^{-1} \\
&= (\mathcal{I}_{g,P}(\theta))^{-1}E_P V\left(h(X;\theta,\alpha)\Omega^{-1}(X;\theta,\beta)g(Y,X;\theta)\,|\,X\right)(\mathcal{I}_{g,P}(\theta))^{-1} \\
&= (\mathcal{I}_{g,P}(\theta))^{-1}E_P\Big[h(X;\theta,\alpha)\Omega^{-1}(X;\theta,\beta)V(g(Y,X;\theta) \mid X) \\
&\quad \times\Omega^{-1}(X;\theta,\beta)h'(X;\theta,\alpha)\Big](\mathcal{I}_{g,P}(\theta))^{-1} \\
&= (\mathcal{I}_{g,P}(\theta))^{-1}.
\end{aligned}$$

Thus, we have established the following property.

Property 23.20: *The modified estimator*

$$\tilde{\theta}_n = \theta_n^* + \left(\sum_{i=1}^{n} h(X_i;\theta_n^*,\alpha_n^*)\Omega^{-1}(X_i;\theta_n^*,\beta_n^*)h'(X_i;\theta_n^*,\alpha_n^*)\right)^{-1} \times \sum_{i=1}^{n} h(X_i;\theta_n^*,\alpha_n^*)\Omega^{-1}(X_i;\theta_n^*,\beta_n^*)g(Y_i,X_i;\theta_n^*),$$

where θ_n^*, α_n^*, and β_n^* are consistent estimators of θ, α, and β, respectively, is an asymptotically efficient estimator in the semiparametric model (23.21).

Remark 23.3: The preceding formula is a direct generalization of the well-known Newton–Raphson formula in the parametric case. In the latter case an identifying constraint is given by the score

$$g(Y,X;\theta) = \frac{\partial \log f(Y \mid X;\theta)}{\partial\theta}.$$

It follows that

$$\begin{aligned} E_P\left(-\frac{\partial g(Y,X;\theta)}{\partial\theta'} \mid X\right) &= -E_P\left(\frac{\partial^2 \log f(Y \mid X;\theta)}{\partial\theta\partial\theta'} \mid X\right) \\ &= V_P\left(\frac{\partial \log f(Y \mid X;\theta)}{\partial\theta} \mid X\right) \\ &= V_P\left(g(Y,X;\theta) \mid X\right). \end{aligned}$$

Thus $h = \Omega$ and the one-step formula simplifies. Specifically, we obtain

$$\begin{aligned} \tilde{\theta}_n &= \theta_n^* + \left(\sum_{i=1}^n h(X_i;\theta_n^*)\right)^{-1} \sum_{i=1}^n g(Y_i,X_i;\theta_n^*) \\ &= \theta_n^* - \left(\sum_{i=1}^n E_{\theta_n^*}\left(\frac{\partial^2 \log f(Y_i \mid X_i;\theta_n^*)}{\partial\theta\partial\theta'} \mid X_i\right)\right)^{-1} \\ &\quad \times \sum_{i=1}^n \frac{\partial \log f(Y_i \mid X_i;\theta_n^*)}{\partial\theta} \\ &\# \quad \theta_n^* - \left(\sum_{i=1}^n \frac{\partial^2 \log f(Y_i \mid X_i;\theta_n^*)}{\partial\theta\partial\theta'}\right)^{-1} \sum_{i=1}^n \frac{\partial \log f(Y_i \mid X_i;\theta_n^*)}{\partial\theta}. \end{aligned}$$

Remark 23.4: In the case of the regression model

$$Y_i = m(X_i,\theta) + u_i,\ E(u_i \mid X_i) = 0,\ V(u_i \mid X_i) = \sigma^2(X_i),$$

a version of the one-step formula is

$$\begin{aligned} \tilde{\theta}_n &= \theta_n^* + \left(\sum_{i=1}^n \frac{1}{\hat{\sigma}_n^2(X_i)} \frac{\partial m(X_i,\theta_n^*)}{\partial\theta} \frac{\partial m(X_i,\theta_n^*)}{\partial\theta'}\right)^{-1} \\ &\quad \times \sum_{i=1}^n \frac{1}{\hat{\sigma}_n^2(X_i)} \frac{\partial m(X_i,\theta_n^*)}{\partial\theta}(Y_i - m(X_i,\theta_n^*)). \end{aligned}$$

23.4 References

Bahadur, R. (1960). "Stochastic Comparison of Tests," *The Annals of Mathematical Statistics*, 31, 276–295.

Bahadur, R. (1967). "Rates of Convergence of Estimates and Tests Statistics," *The Annals of Mathematical Statistics*, 38, 303–324.

Chamberlain, G. (1986). "Asymptotic Efficiency in Semi Parametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

Chamberlain, G. (1987). "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Efron, B. (1975). "Defining the Curvature of a Statistical Problem with Applications to Second-Order Efficiency," *Annals of Statistics*, 3, 1189–1242.

Geweke, J. (1981). "The Approximate Slopes of Econometric Tests," *Econometrica*, 49, 1427–1442.

Gosh, J.K. and Subramanyam, K. (1974). "Second Order Efficiency of Maximum Likelihood Estimators," *Sankhya*, A, 36, 325–358.

Gourieroux, C. (1983). "Asymptotic Comparison of Tests for Non-Nested Hypotheses by Bahadur's A.R.E.," *Proceedings of the 4th Franco–Belgian Meeting of Statisticians*, Publications des Facultés Saint Louis, Bruxelles.

Newey, W.K. (1988). "Adaptative Estimation of Regression Models via Moment Restrictions," *Journal of Econometrics*, 38, 301–339.

Newey, W.K. (1989). "Introduction à la Théorie des Bornes d'Efficacité Semi–paramétriques," *Annales d'Economie et de Statistique*, 13, 1–48.

Pitman, E.J.G. (1979). *Some Basic Theory for Statistical Inference*, Chapman and Hall .

Rao, C.R. (1961). "Asymptotic Efficiency and Limiting Information," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.*

Robinson, P. (1988a). "Semi Parametric Econometrics: A Survey," *Journal of Applied Econometrics*, 3, 35–51.

Robinson, P. (1988b). "Root-N-Consistent Semi Parametric Regression," *Econometrica*, 56, 931–954.

Rothenberg, T.J. (1984). "Approximating the Distributions of Econometric Estimators and Test Statistics," in *Handbook of Econometrics*, Vol.2, North–Holland.

Sargan, J.D. (1976). "Econometric Estimators and the Edgeworth Approximation," *Econometrica*, 44, 421–448.

CHAPTER 24

Asymptotic Theory

In the preceding chapters the main asymptotic properties of estimators and test statistics have often been derived heuristically. In this chapter we present a more rigorous proof of the consistency of these estimators as well as a more detailed derivation of their asymptotic distributions. As a prerequisite we provide some useful tools that underlie the Taylor series expansions that were used to establish, for instance, the asymptotic equivalences among the classical testing procedures.

There are several methods of proofs for establishing the consistency of an "extremum" estimator, i.e., an estimator obtained by maximizing or minimizing a statistical criterion. We have chosen to present those methods that appear to apply to the largest number of situations and for which the regularity conditions are the easiest to verify. Note that the class of extremum estimators contains the M-estimators (maximum likelihood estimator, pseudo maximum likelihood estimators, etc.) as well as moment type estimators (asymptotic least squares estimators, generalized methods of moments, etc.).

24.1 Existence of an Extremum Estimator

Before developing such an asymptotic theory it is necessary to reconsider the definition of an estimator. In Chapter 2 an estimator was defined as a mapping from the set of observations (sample space) to the set of possible values for the parameter (parameter space). In fact, since we are interested in the distribution of an estimator, in its expectation, variance, etc., then this probability distribution must have a meaning, and, for this reason, the mapping should be measurable. In particular, the sample space and the parameter space must be endowed with some σ-algebrae.

Definition 24.1: *Let the sample space Ω and the parameter space Θ be endowed with the σ-algebrae $\mathcal{A}$ and $\mathcal{B}$, respectively. An estimator $\hat{\theta}$ is a measurable mapping from $(\Omega, \mathcal{A})$ to $(\Theta, \mathcal{B})$, i.e., a mapping satisfying*

$$\hat{\theta}^{-1}(B) \in \mathcal{A}, \quad \forall B \in \mathcal{B}.$$

In the preceding chapters we have frequently defined estimators without verifying this measurability requirement. In fact, the measurability of extremum estimators is an immediate consequence of the next property.

Property 24.1: *Suppose that the following assumptions hold:*

A1: *Θ is a compact subset of $\mathbb{R}^p$ with $\mathcal{B} = \mathcal{B}(\Theta)$ being its Borel σ-algebra,*

A2: *The mapping Q from $\Omega \times \Theta$ to $\mathbb{R}$ satisfies:*

(i) $Q(\cdot, \theta) : \omega \to Q(\omega, \theta)$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for every $\theta \in \Theta$,

(ii) the mapping $Q(\omega, \cdot) : \theta \to Q(\omega, \theta)$ is continuous for every $\omega \in \Omega$.

Then there exists a measurable mapping $\hat{\theta}$ from $(\Omega, \mathcal{A})$ to $(\Theta, \mathcal{B}(\Theta))$ such that

$$\forall \omega, \ Q(\omega; \hat{\theta}(\omega)) = \max_{\theta \in \Theta} Q(\omega; \theta).$$

PROOF: The proof consists of four steps.

(i) Recall that every compact subset Θ of $\mathbb{R}^p$ admits a countable dense subset. Thus there exists an increasing sequence of finite subsets Θ_N of Θ such that $\Theta_\infty = \lim_N \uparrow \Theta_N$ is a dense subset of Θ.

(ii) Consider the maximization of Q over the finite subset Θ_N. Let $\theta_1, \ldots, \theta_N$ denote the elements of Θ_N. The function $\hat{\theta}_N$ defined by

$$\begin{aligned}
\hat{\theta}_N &= \theta_1, \ \text{if } Q(\omega; \theta_1) \geq Q(\omega; \theta_j), \ \forall j = 2, \ldots, N, \\
\hat{\theta}_N &= \theta_2, \ \text{if } Q(\omega; \theta_2) > Q(\omega, \theta_1) \\
&\qquad \text{and } Q(\omega, \theta_2) \geq Q(\omega; \theta_j), \ \forall j = 3, \ldots, N, \\
&\vdots \\
\hat{\theta}_N &= \theta_N, \ \text{if } Q(\omega; \theta_N) > Q(\omega; \theta_j), \ \forall j = 1, \ldots, N-1,
\end{aligned}$$

satisfies

$$Q(\omega;\hat{\theta}_N(\omega)) = \max_{\theta\in\Theta_N} Q(\omega;\theta).$$

In addition, the function $\hat{\theta}_N$ is measurable because

$$\{\hat{\theta}_N = \theta_k\} = \bigcap_{j=1}^{k-1}\{Q(\omega,\theta_k) > Q(\omega,\theta_j)\} \bigcap \bigcap_{j=k}^{N}\{Q(\omega,\theta_k) \geq Q(\omega,\theta_j)\},$$

and because all the subsets appearing in the right-hand side are measurable in view of the measurability of $Q(\cdot,\theta)$ for every θ.

(iii) Consider now the function $\hat{\theta}^1(\omega) = \lim_N \inf \hat{\theta}^1_N(\omega)$, where $\hat{\theta}^1_N$ denotes the first component of $\hat{\theta}_N$. Then $\hat{\theta}^1$ is measurable. Moreover, for every value $\omega \in \Omega$, there exists a subsequence $\hat{\theta}_{N_j}(\omega)$ of the sequence $\hat{\theta}_N(\omega)$ that converges to a point

$$(\hat{\theta}^1(\omega), \bar{\theta}^2(\omega), \ldots, \bar{\theta}^p(\omega))'.$$

From the definition of such a subsequence and using the continuity of the function Q, we obtain

$$\begin{aligned}\sup_{(\theta^2,\ldots,\theta^p)} Q(\omega;\hat{\theta}^1(\omega),\theta^2,\ldots,\theta^p) &\geq Q(\omega;\hat{\theta}^1(\omega),\bar{\theta}^2(\omega),\ldots,\bar{\theta}^p(\omega)) \\ &= \lim_{N_j} Q(\omega;\hat{\theta}_{N_j}(\omega)) \\ &= \lim_{N} \sup_{\theta\in\Theta_N} Q(\omega;\theta).\end{aligned}$$

Because Θ_∞ is dense in Θ and because the function Q is continuous, it follows that the function

$$Q^{(1)}(\omega;\theta^2,\ldots,\theta^p) = Q(\omega;\hat{\theta}^1(\omega),\theta^2,\ldots,\theta^p)$$

is a measurable function satisfying

$$\sup_{\theta\in\Theta} Q^{(1)}(\omega;\theta^2,\ldots,\theta^p) \geq \sup_{\theta\in\Theta} Q(\omega;\theta^1,\ldots,\theta^p).$$

(iv) The above argument can be repeated taking $Q^{(1)}$ as the new function. Hence we can construct a measurable function $\hat{\theta}^2$ such that

$$\begin{aligned}\sup_{\theta\in\Theta} Q^{(1)}(\omega;\hat{\theta}^2(\omega),\theta^3,\ldots,\theta^p) &= \sup_{\theta\in\Theta} Q(\omega;\hat{\theta}^1(\omega),\hat{\theta}^2(\omega),\theta^3,\ldots,\theta^p) \\ &\geq \sup_{\theta\in\Theta} Q(\omega;\theta^1,\ldots,\theta^p),\end{aligned}$$

and so on.

Thus, at the p-th step, we have constructed a measurable mapping $\hat{\theta}(\omega) = (\hat{\theta}^1(\omega), \ldots, \hat{\theta}^p(\omega))'$ satisfying

$$Q(\omega; \hat{\theta}(\omega)) = \sup_{\theta \in \Theta} Q(\omega; \theta).$$

□

The asymptotic theory that is developed next is based on the divergence to infinity of the number of observations n. Let the successive observations be denoted $(x_1, y_1), \ldots, (x_i, y_i), \ldots$. Then the sample space Ω is the set of elements $\omega = \{(x_i, y_i), i \in I\!N^*\}$. Now, the objective or criterion function that defines an M-estimator depends on n as well as on ω through $(x_1, y_1), \ldots, (x_n, y_n)$. Such a function is denoted $Q_n(\omega; \theta)$. Hence the M-estimator $\hat{\theta}_n$ is in fact defined as a measurable solution to the optimization problem $\max_{\theta \in \Theta} Q_n(\omega; \theta)$. Then Property 24.1 ensures that such a measurable sequence $\hat{\theta}_n$ exists when Θ is compact and the sequence of objective functions Q_n satisfies assumptions A2-(i) and A2-(ii).

24.2 Consistency of an Extremum Estimator

24.2.1 Almost Sure Consistency

Before giving a criterion ensuring the consistency of an extremum estimator we first recall some definitions concerning statements that hold almost surely.

Let $(\Omega, \mathcal{A}, P_0)$ be a probability space. A property is said to hold P_0-almost surely (P_0 a.s.) if the subset N of ω elements for which the property does not hold is *negligible*, i.e., $P_0(N) = 0$. In particular, a sequence of measurable functions $\hat{\theta}_n$ from $(\Omega, \mathcal{A})$ to $(\Theta, \mathcal{B}(\Theta))$, $\Theta \subset I\!R^p$, *converges P_0-almost surely* to the value θ_0 if and only if

$$P_0\left({}^c\{\limsup_n \hat{\theta}_n(\omega) = \liminf_n \hat{\theta}_n(\omega) = \theta_0\}\right) = 0$$

$$\Leftrightarrow \qquad P_0\left(\lim_n \hat{\theta}_n(\omega) = \theta_0\right) = 1.$$

A sequence of functions Q_n from $\Omega \times \Theta$ to $I\!R$, which is measurable in ω for any given θ, converges uniformly in θ to a function Q_∞, P_0-almost

surely if and only if

$$\sup_{\theta \in \Theta} |Q_n(\cdot;\theta) - Q_\infty(\cdot;\theta)|$$

converges P_0-almost surely to zero.

We note that such a definition implies:

(i) The negligible set N, where the convergence is not uniform, is independent of θ. The set N is said to be *universally negligible.*

(ii) The condition bears upon the supremum over θ of the differences $|Q_n(\cdot;\theta) - Q_\infty(\cdot;\theta)|$.

In the sequel the probability distribution P_0 defined on the sample space is viewed as the true probability distribution.

24.2.2 Some Consistency Results

Property 24.2: *Let P_0 denote the true probability distribution. Suppose that assumptions A1 and A2 (for Q_n) and the following assumptions hold:*

A3: *The sequence of functions Q_n converges uniformly in θ to a function Q_∞, P_0 a.s.,*

A4: *The function Q_∞ is such that the maximum $\max_{\theta \in \Theta} Q_\infty(\omega;\theta)$ is attained at a unique value θ_0, independent of ω.*

Then there exists a sequence of extremum estimators $\hat{\theta}_n$, which are solutions to $\max_{\theta \in \Theta} Q_n(\omega;\theta)$, that converges to θ_0, P_0 a.s.

PROOF: (i) By Property 24.1, assumptions A1 and A2 ensure the existence of a sequence of estimators $\hat{\theta}_n$ satisfying

$$Q_n(\omega;\hat{\theta}_n(\omega)) = \max_{\theta \in \Theta} Q_n(\omega;\theta).$$

(ii) Let N denote the negligible set where Q_n does not converge uniformly to Q_∞. Take an element $\omega \in {}^cN$. Then we can consider the sequence of vectors $\{\hat{\theta}_n(\omega), n \in I\!N\}$. Because this sequence belongs to the compact set Θ, we can find a converging subsequence $\{\hat{\theta}_{n_j(\omega)}(\omega)\}$. Let $\theta_\infty(\omega)$ denote the limit of such a subsequence.

(iii) Because $\lim_j \hat{\theta}_{n_j(\omega)}(\omega) = \theta_\infty(\omega)$, it follows from the uniform convergence and the continuity of the functions $Q_n(\omega, \cdot)$ (and therefore the continuity of $Q_\infty(\omega, \cdot)$) that

$$\lim_j Q_{n_j}(\omega; \hat{\theta}_{n_j(\omega)}(\omega)) = Q_\infty(\omega; \theta_\infty(\omega)).$$

To see this, note that

$$\begin{aligned}
&|Q_{n_j}(\omega; \hat{\theta}_{n_j(\omega)}(\omega)) - Q_\infty(\omega; \theta_\infty(\omega))| \\
&\quad \leq \; |Q_{n_j}(\omega; \hat{\theta}_{n_j}(\omega)) - Q_\infty(\omega; \hat{\theta}_{n_j}(\omega))| \\
&\qquad + |Q_\infty(\omega; \hat{\theta}_{n_j}(\omega)) - Q_\infty(\omega; \theta_\infty(\omega))| \\
&\quad \leq \; \sup_\theta |Q_{n_j}(\omega; \theta) - Q_\infty(\omega; \theta)| + |Q_\infty(\omega; \hat{\theta}_{n_j(\omega)}(\omega)) - Q_\infty(\omega; \theta_\infty(\omega))|.
\end{aligned}$$

(iv) On the other hand, for every value $\theta \in \Theta$, we have

$$Q_{n_j}(\omega; \hat{\theta}_{n_j(\omega)}(\omega)) \geq Q_{n_j}(\omega; \theta).$$

Taking the limits of both sides, we obtain

$$Q_\infty(\omega; \theta_\infty(\omega)) \geq Q_\infty(\omega; \theta), \; \forall \theta \in \Theta.$$

(v) Thus we have $Q_\infty(\omega; \theta_\infty(\omega)) = \max_{\theta \in \Theta} Q_\infty(\omega; \theta)$. Now, because Q_∞ attains its maximum at θ_0 only, then the value $\theta_\infty(\omega)$ must be equal to θ_0.

(vi) Therefore any subsequence of the sequence $\{\hat{\theta}_n(\omega), n \in I\!N\}$ converges to the same value θ_0. Because Θ is compact, it follows that the sequence $\{\hat{\theta}_n(\omega)\}$ itself must converge to θ_0.

(vii) Because this convergence holds for every element $\omega \in {}^cN$, then $\hat{\theta}_n$ converges to θ_0, P_0 a.s., as required. □

24.2.3 Conditions for the Uniform Almost Sure Convergence of the Objective Function

In this subsection we restrict ourselves to M-estimators, i.e., to objective functions of the form

$$Q_n(\omega; \theta) = \frac{1}{n} \sum_{i=1}^{n} \psi(z_i; \theta),$$

where $z_i = (x_i', y_i')'$ is a vector, and $\omega = (z_1, \ldots, z_n, \ldots)$.

Let $P_n(\omega, \cdot)$ denote the empirical probability distribution of the observations $z_1, \ldots, z_n$, i.e., $P_n(\omega, \cdot) = (1/n) \sum_{i=1}^n \epsilon_{z_i}(\cdot)$, where $\epsilon_{z_i}(\cdot)$ denotes the point mass distribution at z_i. Then the objective function can be written under the integral form

$$Q_n(\omega, \theta) = \int_{\mathcal{Z}} \psi(z; \theta) dP_n(\omega; z).$$

a) Finding a Universal Negligible Set

Property 24.3: *Suppose that the empirical distribution $P_n(\omega, .)$ converges in distribution to a limiting distribution $P_\infty(\omega, \cdot)$, P_0 a.s. Then there exists a universal negligible set N such that for any $\omega \in {}^cN$*

$$\lim_n \frac{1}{n} \sum_{i=1}^n \psi(z_i; \theta) = \int_{\mathcal{Z}} \psi(z; \theta) dP_\infty(z) = Q_\infty(\omega, \theta)$$

for every function $\psi(., \theta)$ that is bounded and continuous in z.

PROOF: Take $\omega \in {}^cN$, where N denotes the negligible set where the convergence in distribution does not hold. Then it suffices to note that convergence in distribution is equivalent to convergence of integrals of bounded continuous functions by the Helly–Bray Theorem given in Definition B.32. □

We note that the property of convergence in distribution of the empirical distribution, P_0 a.s., is a classical result of probability theory. It is a direct consequence of the Glivenko–Cantelli Theorem B.9. For instance, we have the following result.

Property 24.4: *Let $\{Z_i, i \in \mathbb{N}\}$ be a sequence of independent random vectors that are identically distributed as P_*. Then $P_n(\omega, \cdot)$ converges in distribution to $P_\infty(\cdot) = P_*(\cdot)$, $P_0 = P_*^{\mathbb{N}}$ a.s.*

In this case we necessarily have

$$Q_\infty(\theta) = \int_{\mathcal{Z}} \psi(z; \theta) dP_*(z).$$

Note that the function Q_∞ is here nonrandom and depends on θ only.

b) Additional Conditions Ensuring Uniform Convergence

It remains to give some conditions ensuring that pointwise convergence, which holds for $\omega \notin N$, also holds uniformly in θ. Below we give three

versions, which are more restrictive than the consistency result of Property 24.2 but which are easier to verify in practice. The first version requires that the objective function is concave.

Property 24.5: *Suppose that the following assumptions hold:*

B1: Θ *is a compact convex subset of* $\mathbb{R}^p$,

B2: *The function* $\psi(z;\theta)$ *is continuous in* θ *and continuous and bounded in* z,

B3: *The function* $\psi(z;\theta)$ *is concave in* θ,

B4: *The empirical distribution* $P_n(\omega,\cdot)$ *converges in distribution to a limiting distribution* $P_\infty(\cdot)$, P_0 *a.s.*,

B5: *The function* $Q_\infty(\theta) = \int_{\mathcal{Z}} \psi(z;\theta) dP_\infty(z)$ *has a unique maximum at* θ_0.

Then there exists a sequence of estimators $\hat{\theta}_n$, *which are solutions to* $\max_{\theta\in\Theta}(1/n)\sum_{i=1}^n \psi(z_i;\theta)$, *that converges to* θ_0, P_0 *a.s.*

PROOF: This property follows at once from Property 24.2, Property 24.3, and the fact that a family $Q_n(\omega,\cdot)$ of concave functions converging pointwise to a function $Q_\infty(\cdot)$, which is necessarily concave, converges uniformly to that function over any compact subset. □

Remark 24.1: For instance, when deriving the properties of the maximum likelihood estimator, i.e., when $\psi(z;\theta) = \log f(z;\theta)$, it may be useful to determine whether the density function is log-concave in the parameters or in some new parameters, which are in a bijective and bicontinuous relationship with the original parameters.

The next version does not require the concavity of the objective function. On the other hand, it requires that ψ is bounded in both of its arguments.

Property 24.6: *Suppose that the following assumptions hold:*

C1: Θ *is a compact subset of* $\mathbb{R}^p$,

C2: *The function* $\psi(z;\theta)$ *is continuous and bounded on* $\mathcal{Z}\times\Theta$,

C3: *The empirical distribution* $P_n(\omega,\cdot)$ *converges in distribution to a limiting distribution* $P_\infty(\cdot)$, P_0 *a.s.*,

C4: *The function* $Q_\infty(\theta) = \int_{\mathcal{Z}} \psi(z;\theta) dP_\infty(z)$ *atttains a unique maximum at* θ_0.

Then there exists a sequence of estimators $\hat{\theta}_n$, *which are solutions to* $\max_{\theta\in\Theta}(1/n)\sum_{i=1}^n \psi(z_i;\theta)$, *that converges to* θ_0, P_0 *a.s.*

PROOF: See Jennrich (1969, Theorem 1). □

Property 24.7: *The same result holds if assumption C2 is replaced by*

C2': *The function* $\psi(z;\theta)$ *is continuous in* θ, *measurable in* z, *and dominated in the sense that*

$$|\sup_{\theta\in\Theta} \psi(z;\theta)| < h(z),$$

where h *is some* P_0*-integrable function.*

PROOF: See Jennrich (1969, Theorem 2). □

When applied to iid observations (see Property 24.4) and to maximum likelihood estimation, then Property 24.7 leads to the following result.

Corollary 24.1: *Let* $Y_1, \ldots, Y_n, \ldots$ *be independently and identically distributed observations with common density* $f(y;\theta_0)$ *with respect to some measure* μ. *Consider the maximum likelihood estimator, which is a solution to*

$$\max_{\theta\in\Theta} \sum_{i=1}^n \log f(y_i;\theta).$$

Suppose that the following assumptions hold:

D1: Θ *is a compact subset of* $\mathbb{R}^p$,

D2: *The density* $f(y;\theta)$ *is continuous in* θ,

D3: *There exists a positive function* $h(y)$ *such that*

$$\left|\log \frac{f(y;\theta)}{f(y;\theta_0)}\right| \leq h(y), \ \forall \in \Theta \text{ and } \int_{\mathcal{Y}} h(y) f(y;\theta_0) d\mu(y) < \infty,$$

D4: θ_0 *is identifiable.*

Then there exists a sequence of maximum likelihood estimators converging almost surely to θ_0.

PROOF: The proof is a direct consequence of the following remarks: (i) every density is measurable in y, (ii) condition D3 is a dominating condition for the function $\psi(y;\theta) = \log(f(y;\theta)/f(y;\theta_0))$, (iii) the limit objective function is $Q_\infty(\theta) = E_\theta \log(f(Y;\theta)/f(Y;\theta_0))$, and (iv) the existence of a unique maximum of Q_∞ is equivalent to the identifiability condition D4. □

The consistency result of Corollary 24.1 is due to Wald (1949).

24.2.4 Consistency of Quasi Generalized Extremum Estimators

The preceding consistency results and in particular Property 24.2 can be easily extended to the case where some of the parameters in the objective function have been replaced by some estimators. Such an extension relies upon the next lemma.

Lemma 24.1: *Let $\tilde{Q}_n(\alpha,\beta)$ be a sequence of functions converging uniformly in (α,β) to $\tilde{Q}_\infty(\alpha,\beta)$. Suppose that the limit function $\tilde{Q}_\infty$ is uniformly continuous in (α,β). Suppose also that $\{\alpha_n\}$ is a sequence converging to α_0. Then the sequence of functions $\tilde{Q}_n(\alpha_n,\beta)$ converges uniformly in β to $\tilde{Q}_\infty(\alpha_0,\beta)$.*

PROOF: We have

$$\begin{aligned}|\tilde{Q}_n(\alpha_n,\beta) - \tilde{Q}_\infty(\alpha_0,\beta)| \;\leq\; & |\tilde{Q}_n(\alpha_n,\beta) - \tilde{Q}_\infty(\alpha_n,\beta)| \\ & + |\tilde{Q}_\infty(\alpha_n,\beta) - \tilde{Q}_\infty(\alpha_0,\beta)|.\end{aligned}$$

Hence

$$\begin{aligned}&\sup_\beta |\tilde{Q}_n(\alpha_n,\beta) - \tilde{Q}_\infty(\alpha_0,\beta)| \\ &\leq \sup_{\alpha,\beta} |\tilde{Q}_n(\alpha,\beta) - \tilde{Q}_\infty(\alpha_n,\beta)| + \sup_\beta |\tilde{Q}_\infty(\alpha_n,\beta) - \tilde{Q}_\infty(\alpha_0,\beta)|.\end{aligned}$$

Because of uniform convergence, the first term on the right-hand side converges to zero as n increases to infinity. Because of uniform continuity, the second term also converges to zero since $\forall\epsilon$, $\exists n_0$ independent of β such that $\|(\alpha_n,\beta)' - (\alpha_0,\beta)'\| < \epsilon$ for every $n \geq n_0$. The desired result follows. □

Property 24.8: *Suppose the assumptions A1, A2, and A3 of Property 24.2 as well as the following assumptions hold:*

A5: *The parameter vector θ can be partitioned as $\theta = (\alpha, \beta)'$, where the two parameter subvectors α and β can vary independently from each other, i.e., $(\alpha, \beta) \in A \times B$,*

A6: *$\tilde{\alpha}_n$ is an estimator converging P_0 a.s. to some value α_0,*

A7: *The function $\tilde{Q}_\infty$ is such that the maximum $\max_{\beta \in B} \tilde{Q}_\infty(\omega; \alpha_0, \beta)$ is attained at a unique value β_0.*

Then there exists a sequence of quasi generalized extremum estimators $\tilde{\beta}_n$, which are solutions to

$$\max_{\beta \in B} \tilde{Q}_n(\omega; \tilde{\alpha}_n(\omega), \beta),$$

that converges to β_0, P_0 a.s.

PROOF: Consider the function

$$Q_n(\omega; \beta) = \tilde{Q}_n(\omega; \tilde{\alpha}_n(\omega), \beta).$$

Then it suffices to verify that assumptions A1, A2, A3, and A4 of Property 24.2 are satisfied:

(i) Assumption A1 follows from the fact that the projection of a compact subset is a compact subset.

(ii) Assumption A2 is trivially satisfied.

(iii) The uniform convergence of $\tilde{Q}_n(\omega; \beta)$ (given ω) follows from assumption A3 on the uniform convergence of $\tilde{Q}_n(\omega; \theta)$, the fact that the limit of $\tilde{Q}_\infty(\omega; \theta)$, which is continuous on the compact set Θ, is uniformly continuous, and Lemma 24.1.

(iv) Lastly, assumption A4 follows from assumption A7.

□

Thus the regularity conditions ensuring the consistency of a quasi generalized extremum estimator of β are similar to those ensuring the consistency of the extremum estimator of the pair $\theta = (\alpha, \beta)'$.

24.3 First-Order Conditions

24.3.1 Unconstrained Extremum Estimators

Property 24.9: *Let $\{\hat{\theta}_n\}$ be a sequence of extremum estimators of θ obtained by maximizing the objective function $Q_n(\omega;\theta)$. Suppose that the sequence $\hat{\theta}_n$ converges P_0 a.s. to a point θ_0 belonging to the interior of the parameter space Θ. Suppose also that the function $Q_n(\omega;\cdot)$ is differentiable with respect to θ. Then the estimator $\hat{\theta}_n$ satisfies asymptotically and P_0 a.s. the first-order conditions*

$$\frac{\partial Q_n(\omega;\hat{\theta}_n(\omega))}{\partial\theta} = 0.$$

PROOF: Let N denote the negligible subset where the above P_0 a.s. convergence does not hold.

Consider a neighborhood $\mathcal{V}$ of θ_0 included in Θ. For every element $\omega \in {}^cN$, there exists a minimum integer $n(\omega)$ such that $\hat{\theta}_n(\omega)$ belongs to $\mathcal{V}$ for all $n > n(\omega)$. Because $\hat{\theta}_n(\omega)$ corresponds to a maximum and because this maximum is attained on the interior of Θ, it follows that $\hat{\theta}_n(\omega)$ must satisfy the corresponding first-order conditions. That is, we have

$$\forall\, \omega \in {}^cN, \exists\, n(\omega) : \forall n \geq n(\omega) : \frac{\partial Q_n(\omega;\hat{\theta}_n(\omega))}{\partial\theta} = 0.$$

We now make clear the meaning of the qualification "asymptotically" in the above statement. We note that the preceding first-order conditions are satisfied when $n > n(\omega)$, the latter number being random. Because this minimal number is finite for every $\omega \in {}^cN$, we have the following property

$$\lim_{n_0\to\infty} P_0\left(\frac{\partial Q_n(\omega;\hat{\theta}_n(\omega))}{\partial\theta} = 0, \forall\, n \geq n_0\right) = \lim_{n_0\to\infty} P_0(n(\omega) \leq n_0) = 1.$$

□

Clearly, Property 24.9 can be generalized to quasi generalized extremum estimators.

24.3.2 Constrained Extremum Estimators

When considering estimation under constraints, a main difference is that the parameter space Θ is now of empty interior. However, it may frequently be written as

$$\Theta = \Theta^* \cap \{\theta : g(\theta) = 0\},$$

where Θ^* is of nonempty interior. This is assumed hereafter. It is also assumed that none of the r constraints in $g(\theta) = 0$ is redundant.

Then a proof similar to the proof of Property 24.9 gives the following result.

Property 24.10: *Let $\{\hat{\theta}_n\}$ be a sequence of extremum estimators, which are solutions to $\max_{\theta \in \Theta} Q_n(\omega, \theta)$, i.e., which are solutions to the constrained optimizations problems*

$$\begin{cases} \max_{\theta \in \Theta^*} Q_n(\omega, \theta) \\ \text{subject to } g(\theta) = 0. \end{cases}$$

Assume that the sequence $\{\hat{\theta}_n\}$ converges P_0 a.s. to a point θ_0 in Θ belonging to the interior of Θ^*. Assume also that the functions Q_n and g are differentiable in θ. Then the estimator $\hat{\theta}_n$ satisfies asymptotically and P_0 a.s. the first-order conditions

$$\frac{\partial Q_n(\omega; \hat{\theta}_n(\omega))}{\partial \theta} - \frac{\partial g'(\hat{\theta}_n(\omega))}{\partial \theta} \hat{\lambda}_n(\omega) = 0,$$
$$g(\hat{\theta}_n(\omega)) = 0,$$

where $\hat{\lambda}_n(\omega)$ is a sequence of Lagrange multipliers.

When considering constrained estimation, it is useful to study the properties of the sequence of Lagrange multipliers $\hat{\lambda}_n(\omega)$.

Property 24.11: *Under the assumptions of Property 24.10, if the Jacobian matrix $(\partial g'(\theta)/\partial \theta)$ is measurable with full column rank r on Θ, then the Lagrange multipliers $\hat{\lambda}_n(\omega)$ exist asymptotically P_0 a.s., and are measurable functions.*

PROOF: Let E_n denote the subset of Ω where the first-order conditions in Property 24.10 are satisfied for a sample of size larger than n. We know that $\lim_{n \to \infty} P_0(E_n) = 1$ and that the sequence E_n is increasing.

Then, when ω belongs to E_n, the Lagrange multipliers are well-defined. Specifically, we have

$$\hat{\lambda}_n(\omega) = \left(\frac{\partial g(\hat{\theta}_n(\omega))}{\partial \theta'} \frac{\partial g(\hat{\theta}_n(\omega))}{\partial \theta'} \right)^{-1} \frac{\partial g(\hat{\theta}_n(\omega))}{\partial \theta'} \frac{\partial Q_n(\omega; \hat{\theta}_n(\omega))}{\partial \theta}.$$

This establishes the existence and the measurability of the Lagrange multipliers on the set E_n. Therefore the Lagrange multipliers $\hat{\lambda}_n$ exist asymptotically and P_0 a.s., i.e., on a set E_n whose probability converges to one.

If we now want to define the Lagrange multipliers in every finite sample, it suffices to extend the preceding definition of $\hat{\lambda}_n$, while retaining the measurability property. For instance, we can set the Lagrange multipliers to be equal to some arbitrary values such as zero when ω belongs to ${}^cE_n$. □

Property 24.12: *Suppose that the assumptions of Property 24.11 are satisfied. Suppose also that:*

(i) the function g is continuously differentiable in θ,

(ii) the sequence of functions $\partial Q_n(\omega;\theta)/\partial\theta$ converges uniformly in θ, P_0 a.s., to a continuous function $\dot{Q}_\infty(\theta)$ as n increases to infinity.

Then the sequence of Lagrange multipliers $\hat{\lambda}_n$ converges P_0 a.s. to

$$\lambda_\infty = \left(\frac{\partial g(\theta_0)}{\partial \theta'} \frac{\partial g'(\theta_0)}{\partial \theta} \right)^{-1} \frac{\partial g(\theta_0)}{\partial \theta'} \dot{Q}_\infty(\theta_0).$$

PROOF: It suffices essentially to permute the order of the limits, which is justified in view of our continuity assumptions. Specifically, let $\bar{N}$ denote the negligible set outside which the preceding P_0 a.s. convergences do not hold.

If $\omega \in {}^c\bar{N}$, then we have

$$\lim_{n\to\infty} \frac{\partial g(\hat{\theta}_n(\omega))}{\partial \theta'} = \frac{\partial g(\theta_0)}{\partial \theta'},$$

where we have used the continuity of the Jacobian. Moreover, using assumption (ii) and Lemma 24.1, we have

$$\lim_{n\to\infty} \frac{\partial Q_n(\omega; \hat{\theta}_n(\omega))}{\partial \theta} = \dot{Q}_\infty(\theta_0).$$

It follows that, if $\omega \in {}^{c}\bar{N}$, then the Lagrange multipliers converge to

$$\left(\frac{\partial g(\theta_0)}{\partial \theta'}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}\frac{\partial g(\theta_0)}{\partial \theta'}\dot{Q}_\infty(\theta_0).$$

□

Corollary 24.2: *In addition to the preceding assumptions, suppose that:*

(i) the function Q_∞ is differentiable in θ,

(ii) $\dot{Q}_\infty(\theta) = \partial Q_\infty(\theta)/\partial\theta$.

Then the sequence of Lagrange multipliers $\hat{\lambda}_n$ converges P_0 a.s. to zero.

PROOF: Note that $\theta_0 = \lim \hat{\theta}_n$ is a value at which Q_∞ attains its maximum. Because θ_0 belongs to the interior of Θ and because Q_∞ is differentiable, it follows that

$$\frac{\partial Q_\infty(\theta_0)}{\partial \theta} = 0.$$

Thus we have

$$\lambda_\infty = \left(\frac{\partial g(\theta_0)}{\partial \theta'}\frac{\partial g'(\theta_0)}{\partial \theta}\right)^{-1}\frac{\partial g(\theta_0)}{\partial \theta'}\dot{Q}_\infty(\theta_0) = 0,$$

which establishes the desired result. □

Note that condition (ii) can be viewed as allowing us to permute a limit operator and a differentiation operator, namely

$$\lim_{n\to\infty}\left(\frac{\partial Q_n(\omega;\theta)}{\partial \theta}\right) = \frac{\partial\left(\lim_{n\to\infty} Q_n(\omega;\theta)\right)}{\partial \theta}.$$

24.4 Taylor Series Expansions

Lastly, we shall give a justification of the Taylor series expansions of the first-order conditions that were used to derive, for instance, the asymptotic distributions of extremum estimators and the asymptotic equivalences among the main classical testing procedures. It is first necessary to introduce the concept of quantities that are infinitely small in probability. We shall also need to establish a Mean Value Theorem for random variables.

24.4.1 Infinitely Small Quantities in Probability

Definition 24.2: *A sequence of random vectors* $\{X_n, n \in \mathbb{N}\}$ *defined on the probability space* $(\Omega, \mathcal{A}, P_0)$ *is infinitely small in probability if* X_n *converges in probability to zero. Such a property is denoted by* $X_n = o_p(1)$.

From Definition B.29 of convergence in probability it follows that

$$X_n = o_p(1) \Leftrightarrow \forall \epsilon > 0, \lim_{n \to \infty} P_0(|X_n| > \epsilon) = 0.$$

Because convergence in probability is closed under any continuous transformation (see Property B.68), and because the convergence in probability of a random vector is equivalent to the convergence in probability of each component of this vector, then we obtain at once the following property.

Property 24.13: *Let* $\{Y_n, n \in \mathbb{N}\}$ *be a sequence of random vectors converging in probability to* Y_∞. *Let* $\{X_n, n \in \mathbb{N}\}$ *be an infinitely small quantity in probability. Let* f *be a continuous function of* (X, Y) *such that* $f(0, Y_\infty) = 0$. *Then the sequence* $\{f(X_n, Y_n)\}$ *is an infinitely small quantity in probability.*

Applying Property 24.13 to appropriate functions f, we obtain the following implications

$$X_n = o_p(1), Y_n = o_p(1) \Rightarrow X_n + Y_n = o_p(1), \tag{24.1}$$

assuming that the vectors have the same dimension

$$X_n = o_p(1), Y_n \xrightarrow{P_0} Y_\infty \Rightarrow X_n Y_n = o_p(1), \tag{24.2}$$

assuming that the vectors or matrices X_n and Y_n are compatible.

Other useful properties relating the concept of infinitely small quantities in probability to the concept of convergence in distribution can be established. Such properties can be proved using characteristic functions. We have, however, preferred to give some proofs relying on more elementary concepts.

Property 24.14: *If* $X_n = o_p(1)$ *and if* Y_n *converges in distribution to a random variable* Y, *then* $X_n + Y_n$ *converges in distribution to* Y.

PROOF: For every pair of events (E, A) we have the equality

$$P_0(E) \leq P_0(E \cap A) + P_0({}^cA).$$

It now suffices to apply this equality to appropriately chosen pairs of events.

(i) Take $E_n = \{X_n + Y_n \leq y\}$ and $A_n = \{\|X_n\| \leq \epsilon\}$. Then there exists n_ϵ such that

$$\forall n \geq n_\epsilon, P({}^cA_n) \leq \epsilon.$$

Hence we have

$$\begin{aligned} P(X_n + Y_n \leq y) &\leq P[(X_n + Y_n \leq y) \cap (|X_n| \leq \epsilon)] + \epsilon \\ &\leq P(Y_n \leq y + \epsilon) + \epsilon. \end{aligned}$$

(ii) Suppose now that we take $E_n = \{Y_n \leq y - \epsilon\}$. Then we have

$$\begin{aligned} P(Y_n \leq y - \epsilon) &\leq P[(Y_n \leq y - \epsilon) \cap (|X_n| \leq \epsilon)] + \epsilon \\ &\leq P(X_n + Y_n \leq y) + \epsilon. \end{aligned}$$

(iii) Collecting results, we have

$$P(Y_n \leq y - \epsilon) - \epsilon \leq P(X_n + Y_n \leq y) \leq P(Y_n \leq y + \epsilon) + \epsilon.$$

Now, if $y-\epsilon$ and $y+\epsilon$ are continuity points of the limiting distribution of Y_n, then we obtain

$$\begin{aligned} G(y - \epsilon) - \epsilon &\leq \lim_n \inf P(X_n + Y_n \leq y) \\ &\leq \lim_n \sup P(X_n + Y_n \leq y) \\ &\leq G(y + \epsilon) + \epsilon, \end{aligned}$$

where G denotes the cumulative distribution function of the limiting distribution of Y_n. These inequalities hold for every $\epsilon > 0$, where $y - \epsilon$ and $y + \epsilon$ are continuity points of G.

If y is a continuity point of G, then we can find a sequence ϵ_k converging to zero such that $y - \epsilon_k$ and $y + \epsilon_k$ are also continuity points. Thus, taking the limit, we obtain

$$G(y) \leq \lim_n \inf P(X_n + Y_n \leq y) \leq \lim_n \sup P(X_n + Y_n \leq y) \leq G(y).$$

Hence we have

$$\lim_{n \to \infty} P(X_n + Y_n \leq y) = G(y),$$

for every continuity point y of G. □

Property 24.15: *If $X_n = o_p(1)$ and if Y_n converges in distribution to a random variable Y, where X_n and Y_n are scalars, then their product satisfies $X_n Y_n = o_p(1)$.*

PROOF: We have

$$\begin{aligned} P(|X_nY_n| > \epsilon) &\leq P(|X_nY_n| > \epsilon, |Y_n| \leq M) + P(|Y_n| > M) \\ &\leq P\left(|X_n| > \frac{\epsilon}{M}\right) + P(|Y_n| > M). \end{aligned}$$

Letting n increase to infinity, it follows that, if M and $-M$ are continuity points of the limiting cumulative distribution function of Y_n, then

$$\begin{aligned} \lim_n \sup P(|X_nY_n| > \epsilon) &\leq \lim_n P\left(|X_n| > \frac{\epsilon}{M}\right) + \lim_n P(|Y_n| > M) \\ &= 0 + 1 - G(M) + G(-M). \end{aligned}$$

It now suffices to let M increase to $+\infty$. Then we obtain

$$\lim_n \sup P(|X_nY_n| > \epsilon) \leq \lim_M (1 - G(M) + G(-M)) = 0,$$

which establishes the desired result. □

Lastly, we shall frequently use a Mean Value Theorem for random variables.

Mean Value Theorem 24.1: *Let q be a function from $\Omega \times \Theta$ with values in $\mathbb{R}$ that is measurable in ω for every θ and continuously differentiable in θ for every ω. Suppose that Θ is compact and convex. If $\hat{\theta}_1(\omega)$ and $\hat{\theta}_2(\omega)$ are two measurable functions from $(\Omega, \mathcal{A})$ to $(\Theta, \mathcal{B}(\Theta))$, then there exists a measurable function $\bar{\theta}(\omega)$ defined on the same spaces such that:*

(i)

$$q(\omega; \hat{\theta}_1(\omega)) - q(\omega; \hat{\theta}_2(\omega)) = \frac{\partial q(\omega; \bar{\theta}(\omega))}{\partial \theta'}(\hat{\theta}_1(\omega) - \hat{\theta}_2(\omega)),$$

(ii) $\bar{\theta}(\omega)$ belongs to the segment $[\hat{\theta}_1(\omega), \hat{\theta}_2(\omega)]$ for every $\omega \in \Omega$.

PROOF: It suffices to establish the measurability of the function $\bar{\theta}$ defining the mean value $\bar{\theta}(\omega)$. Let

$$\begin{aligned} Q(\omega; \theta) &= \left| q(\omega; \hat{\theta}_1(\omega)) - q(\omega; \hat{\theta}_2(\omega)) - \frac{\partial q(\omega; \theta)}{\partial \theta'}(\hat{\theta}_1(\omega) - \hat{\theta}_2(\omega)) \right| \\ &\quad + D(\omega; \theta), \end{aligned}$$

where $D(\omega; \theta)$ denotes the distance between θ and the segment defined by $\hat{\theta}_1(\omega)$ and $\hat{\theta}_2(\omega)$.

From Property 24.1 we know that there exists a measurable solution $\bar{\theta}(\omega)$ (say) to the minimization problem $\min_{\theta\in\Theta} Q(\omega;\theta)$.

On the other hand, it follows from the usual Mean Value Theorem applied to every fixed ω that the above minimum is zero. Hence we have simultaneously

$$q(\omega;\hat{\theta}_1(\omega)) - q(\omega;\hat{\theta}_2(\omega)) - \frac{\partial q(\omega;\bar{\theta}(\omega))}{\partial\theta'}(\hat{\theta}_1(\omega) - \hat{\theta}_2(\omega)) = 0$$

and

$$D(\omega;\bar{\theta}(\omega)) = 0,$$

i.e. conditions (i) and (ii), as required. □

24.4.2 Taylor Series Expansion of the First-Order Conditions and Asymptotic Normality

We assume that the assumptions of Property 24.2 are satisfied, i.e., existence and P_0 a.s. consistency of the estimator $\hat{\theta}_n$, as well as the assumptions of Property 24.9, which justify the use of the first-order conditions. In addition, we make the following assumptions:

E1: The objective function Q_n is twice continuously differentiable in θ,

E2: The Hessian $\left(\partial^2 Q_n(\omega;\theta)/\partial\theta\partial\theta'\right)$ converges uniformly to the limit function $\ddot{Q}_\infty(\theta_0)$, P_0 a.s.,

E3: The matrix $\ddot{Q}_\infty(\theta_0)$ is nonsingular,

E4: The sequence $\{\sqrt{n}(\partial Q_n(\omega;\theta_0)/\partial\theta)\}$ converges in distribution to the multivariate normal distribution $N(0,\mathcal{J}(\theta_0))$.

From the P_0 a.s. consistency of the extremum estimator, Property 24.9, and assumptions E1,E2, and E3, it follows that we can find an increasing sequence of subsets $\tilde{E}_n$ of Ω such that $\lim_{n\to\infty} P_0(\tilde{E}_n) = 1$ and such that, for every $\omega \in \tilde{E}_n$, we have simultaneously:

(i) $\partial Q_n(\omega;\hat{\theta}_n(\omega))/\partial\theta = 0$,

(ii) if $\bar{\theta}_{jn}(\omega), j = 1,\ldots,p$, are arbitrary points in the segment defined by $\hat{\theta}_n(\omega)$ and θ_0, then the square matrix of size p with j-th row $\partial^2 Q_n(\omega;\bar{\theta}_{jn}(\omega))/\partial\theta_j\partial\theta'$ is nonsingular. See assumption E2 and E3.

Before applying our preceding results on infinitely small quantities, it is first necessary to extend the system of first-order conditions so as to define it for every $\omega \in \Omega$. In the sequel we let

$$\dot{Q}_n(\omega;\theta) = \begin{cases} \dfrac{\partial Q_n(\omega;\theta)}{\partial\theta} & , \text{ if } \omega \in \tilde{E}_n, \\ 0 & , \text{ if } \omega \notin \tilde{E}_n, \end{cases}$$

$$\tilde{\theta}_n(\omega) = \begin{cases} \hat{\theta}_n(\omega) & , \text{ if } \omega \in \tilde{E}_n, \\ \theta_0 & , \text{ if } \omega \notin \tilde{E}_n. \end{cases}$$

Thus we can write

$$\forall \omega \in \Omega,\ \dot{Q}_n(\omega;\tilde{\theta}_n(\omega)) = 0.$$

Note that our extension implies that $\tilde{\theta}_n$ is P_0 a.s. consistent for θ_0.

We can now consider a Taylor series expansion of the first-order conditions $\dot{Q}_n(\omega;\tilde{\theta}_n(\omega)) = 0$ using the Mean Value Theorem 24.1. Such an expansion has to be carried for each component of $\dot{Q}_n$. Let $\bar{\theta}_{jn}(\omega)$ denote a value in the segment $[\tilde{\theta}_n(\omega), \theta_0]$. Then we have

$$\begin{aligned} \dot{Q}_{jn}(\omega;\tilde{\theta}_n(\omega)) &= \dot{Q}_{jn}(\omega;\theta_0) + \frac{\partial \dot{Q}_{jn}(\omega;\bar{\theta}_{jn}(\omega))}{\partial\theta'}(\tilde{\theta}_n(\omega) - \theta_0) \\ &= 0, \quad j = 1,\ldots,p \end{aligned}$$

Distinguishing the case where ω belongs to $\tilde{E}_n$ from the case where it does not, we see that

$$\text{if } \omega \in \tilde{E}_n : \frac{\partial Q_n(\omega;\theta_0)}{\partial\theta_j} = -\frac{\partial^2 Q_n(\omega;\bar{\theta}_{jn}(\omega))}{\partial\theta_j\partial\theta'}(\tilde{\theta}_n(\omega) - \theta_0)$$

$$\text{if } \omega \notin \tilde{E}_n : 0 = -\ddot{Q}_\infty(\theta_0)(\tilde{\theta}_n(\omega) - \theta_0),$$

i.e.

$$\text{if } \omega \in \tilde{E}_n : \sqrt{n}(\tilde{\theta}_n(\omega) - \theta_0) = -\begin{pmatrix} \dfrac{\partial^2 Q_n(\omega;\bar{\theta}_{1n}(\omega))}{\partial\theta_1\partial\theta'} \\ \vdots \\ \dfrac{\partial^2 Q_n(\omega;\bar{\theta}_{pn}(\omega))}{\partial\theta_p\partial\theta'} \end{pmatrix}^{-1} \sqrt{n}\frac{\partial Q_n(\omega;\theta_0)}{\partial\theta},$$

$$\text{if } \omega \notin \tilde{E}_n : \sqrt{n}(\tilde{\theta}_n(\omega) - \theta_0) = -(\ddot{Q}_\infty(\theta_0))^{-1} \times 0.$$

Because the Hessian is continuous (assumption E1), the convergence of the Hessian is uniform (assumption E2), and the mean values $\bar{\theta}_{jn}$, $j =$

$1, \ldots, p$, converge P_0 a.s. to θ_0, then it follows that

$$\forall \omega \in \Omega : \sqrt{n}(\tilde{\theta}_n(\omega) - \theta_0) = \left((-\ddot{Q}_\infty(\theta_0))^{-1} + o_p(1)\right) \sqrt{n}\dot{Q}_n(\omega; \theta_0).$$

But $\sqrt{n}\dot{Q}_n(\omega; \theta_0) = \sqrt{n}(\partial Q_n(\omega; \theta_0)/\partial\theta)$ when $\omega \in \tilde{E}_n$. Hence we have

$$P\left(\sqrt{n}\dot{Q}_n(\omega; \theta_0) \neq \sqrt{n}\frac{\partial Q_n(\omega; \theta_0)}{\partial\theta}\right) \leq P({}^c\tilde{E}_n),$$

which converges to zero as n increases to infinity. Thus we can write

$$\sqrt{n}\dot{Q}_n(\omega; \theta_0) = \sqrt{n}\frac{\partial Q_n(\omega; \theta_0)}{\partial\theta} + o_p(1).$$

Then, using property 24.14, we obtain the following convergence in distribution

$$\sqrt{n}\dot{Q}_n(\omega; \theta_0) \xrightarrow{d} N(0, \mathcal{J}(\theta_0)).$$

Next, applying Property 24.15, we obtain

$$\sqrt{n}(\tilde{\theta}_n(\omega) - \theta_0) = (-\ddot{Q}_\infty(\theta_0))^{-1}\sqrt{n}\frac{\partial Q_n(\omega; \theta_0)}{\partial\theta} + o_p(1).$$

Because $\tilde{\theta}_n$ differs from $\hat{\theta}_n$ on ${}^c\tilde{E}_n$ only, it follows that we also have

$$\sqrt{n}(\hat{\theta}_n(\omega) - \theta_0) = -(\ddot{Q}_\infty(\theta_0))^{-1}\sqrt{n}\frac{\partial Q_n(\omega; \theta_0)}{\partial\theta} + o_p(1). \qquad (24.3)$$

We note that the preceding expansion is valid irrespective of the definition chosen for $\hat{\theta}_n$ outside of the set where $\hat{\theta}_n$ is well-defined. In particular, from Property 24.14 and assumption E4, the vectors $\sqrt{n}(\hat{\theta}_n(\omega) - \theta_0)$ and $-(\ddot{Q}_\infty(\theta_0))^{-1}\sqrt{n}(\partial Q_n(\omega; \theta_0)/\partial\theta)$ converge in distribution to the same random variable. Thus we have proved the following property.

Property 24.16: *Under the assumptions of Properties 24.2 and 24.9, and assumptions E1–E4, we have*

$$\sqrt{n}(\hat{\theta}_n(\omega) - \theta_0) \xrightarrow{d} N(0, (\ddot{Q}_\infty(\theta_0))^{-1}\mathcal{J}(\theta_0)(\ddot{Q}_\infty(\theta_0))^{-1}).$$

24.5 References

Gallant, A.R. (1987). *Non Linear Statistical Models,* Wiley.

Gallant, A.R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Non Linear Dynamic Models,* Basil Blackwell.

Jennrich, R. (1969). "Asymptotic Properties of Non Linear Least Squares Estimators, " *The Annals of Mathematical Statistics*, 40, 633-643.

Malinvaud, E. (1970). "The Consistency of Non Linear Regressions," *The Annals of Mathematical Statistics*, 41, 956-969.

Wald, J. (1949). "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics*, 20, 595-601.

White, H. (1934). *Asymptotic Theory for Econometricians*, Academic Press.

APPENDIX A

Review of Linear Algebra and Matrix Calculus

A.1 Linear Algebra and Notations

A.1.1 Vector Spaces

A *vector space* over the field $\mathbb{R}$ is a set E endowed with two operations, which are the addition and the multiplication. These operations must satisfy the following properties:

(i) **Properties of the addition:** The addition, denoted "+", is a binary operation defined on E, i.e., such that $\forall x, y \in E, x+y \in E$, that gives a commutative group structure to the set E.

(ii) **Properties of the multiplication:** The multiplication satisfies $\forall x \in E, \forall a \in \mathbb{R} : ax \in E$. Moreover, $1x = x$.

(iii) **Properties involving the addition and the multiplication**

$$\forall x, y \in E,\ \forall a, b \in \mathbb{R}: \qquad \begin{aligned} a(x+y) &= ax + ay, \\ (a+b)x &= ax + bx. \end{aligned}$$

The elements of a vector space are called *vectors*. In particular, the preceding properties imply the existence of a vector 0 satisfying $x + 0 = 0 + x = 0$. This vector is called the *zero vector*. Given a finite set or *system* of vectors $x_1, \ldots, x_n$, the preceding properties also allows us to consider vectors of the forms $x = a_1x_1 + \ldots + a_nx_n$, where $a_1, \ldots, a_n \in \mathbb{R}$. Such a vector x is said to be *a linear combination* of $x_1, \ldots, x_n$.

A vector space F included in the vector space E is called a *subspace* of E. It is characterized by the property that every linear combination

of any two of its elements is one of its elements. Note that the smallest subspace containing a given system of vectors $x_1, \ldots, x_n$ always exists. It is obtained by taking all possible linear combinations of the vectors of this system. Such a subspace is said to be *generated* or *spanned* by $x_1, \ldots, x_n$.

A.1.2 Linear Mappings

Given two vector spaces E and F over $\mathbb{R}$, a mapping f from E to F is *linear* if

$$\forall x, y \in E,\ \forall a, b \in \mathbb{R} : f(ax + by) = af(x) + bf(y).$$

The set of linear mappings from E to F is itself a vector space, which is denoted $\mathcal{L}(E, F)$. In the special case where $F = \mathbb{R}$ ($\mathbb{R}$ is clearly a vector space), then f is called a *linear form* instead of a linear mapping. Moreover, the set $\mathcal{L}(E, \mathbb{R})$ is called *the dual space* of E and is denoted E^*.

Let E, F, and G be three vector spaces. If f is a linear mapping from E to F, and if g is a linear mapping from F to G, then their composition, denoted $h = g \circ f$, is a linear mapping from E to G. The composition is associative but is not commutative. Thus, in general, $g \circ f \neq f \circ g$.

The direct image and inverse image of a subspace by a linear mapping is still a subspace. In particular, if f is a linear mapping from E to F, then $f(E)$ is a subspace of F, called the *image* or *span* of f. It is denoted Im f. By the same property, $f^{-1}(0) = \{x \in E : f(x) = 0\}$ is a subspace of E, called the *kernel* or *null space* of f. It is denoted ker f.

The knowledge of these two particular subspaces is important because of the following properties:

(i) f is *injective* or one-to-one if ker $f = \{0\}$,

(ii) f is *surjective* or onto if Im $f = F$.

When f is simultaneously injective and surjective, it is said to be *bijective*. Then its inverse denoted f^{-1} exists. Moreover, the inverse is a linear mapping from F to E.

A.1.3 Bases

A system of vectors $x_1, \ldots, x_n$ of E is said to be *linearly independent* if and only if

$$a_1 x_1 + \ldots + a_n x_n = 0 \Rightarrow a_1 = \ldots = a_n = 0.$$

A system $x_1, \ldots, x_n$ is said to be a *generating system* of E if every vector of E can be written as a linear combination of the vectors of such a system. A *basis* of E is simultaneously a linearly independent and generating system of E. A vector space possesses many bases. If one of these bases is made of n elements, then any other basis has the same number n of elements. Thus the number n is a characteristic specific to the vector space. The vector space E is said to be *finite-dimensional* with dimension $\dim E$ equal to n if n is finite. In this case a vector x of E can be uniquely written as a linear combination $x = a_1x_1 + \ldots + a_nx_n$ of the elements of the basis, i.e., the *basis vectors* $x_1, \ldots, x_n$. The real numbers $a_1, \ldots, a_n$ are called the *coordinates* or *components* of x with respect to this basis.

The dimension of a subspace F of E is always smaller or equal to the dimension of E. When $\dim F = \dim E$, then F is equal to E. When $\dim F = 0$, then $F = \{0\}$.

The conditions given in Section A.1.2 for a linear mapping to be either injective or surjective can be written as conditions on the dimensions of some appropriate spaces. Namely, let f be a linear mapping from the vector space E to the vector space F. The *rank* of f is the dimension of the subspace $\mathrm{Im} f$. Hence, by definition, we have $\mathrm{rank} f = \dim \mathrm{Im}\ f = \dim f(E)$. Moreover, we have

$$\mathrm{rank} f + \dim \ker f = \dim E.$$

In particular, if the vector spaces E and F have the same dimension, then the linear mapping f is bijective if and only if it is injective, i.e., if and only if $\ker f = \{0\}$.

A.1.4 Supplementary Subspaces

Consider two subspaces F_1 and F_2 of a vector space E. The *sum* of F_1 and F_2 is the set denoted $F_1 + F_2$ of vectors of the form $x_1 + x_2$, where $x_1 \in F_1$ and $x_2 \in F_2$.

It is easy to verify that $F_1 + F_2$ is a subspace of E. If in addition $F_1 \cap F_2 = \{0\}$, then the sum $F_1 + F_2$ is said to be the *direct sum* of the subspaces F_1 and F_2. This direct sum is denoted $F_1 \oplus F_2$. Now, the sum of F_1 and F_2 is direct if and only if every vector x of $F_1 + F_2$ can be written as $x = x_1 + x_2$ for some unique $x_1 \in F_1$ and some unique $x_2 \in F_2$.

When $F_1 \oplus F_2 = E$, then F_1 and F_2 are said to be two *supplementary spaces* of E.

It can be shown that two subspaces F_1 and F_2 of a vector space E of finite dimension are supplementary if and only if the following two conditions are satisfied:

(i) $\dim F_1 + \dim F_2 = n$,

(ii) $F_1 \cap F_2 = \{0\}$.

When F_1 and F_2 are supplementary subspaces, then a basis of E can be obtained by taking the union of a basis of F_1 and a basis of F_2.

A.1.5 Matrix Representations

Throughout this subsection we consider two vector spaces E and F of dimension p and n, respectively. Let $(u_1, \ldots, u_p)$ be a basis of E and $(v_1, \ldots, v_n)$ be a basis of F.

A vector x of F can be written as a unique combination $x = \sum_{i=1}^{n} a_i v_i$ of the elements of the basis of F, where $a_i \in \mathbb{R}, i = 1, \ldots, n$. Thus the vector x is fully characterized by the *column vector* of its components, namely $x = (a_1, a_2, \ldots, a_n)'$, where "$'$" denotes transposition (see Section A.1.6).

Such a representation can be extended to describe a system of vectors of F. Let $x_1, \ldots, x_j$ be such a system. Each vector of this system can be uniquely written as $x_j = \sum_{i=1}^{n} a_{ij} v_i$, $j = 1, \ldots, J$. Hence the system of vectors is fully characterized by the knowledge of

$$\mathbf{A} = \begin{pmatrix} a_{11} & \ldots & a_{1J} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nJ} \end{pmatrix}.$$

Such a table is called the *matrix representation of the system* $x_1, \ldots, x_J$. It consists of n *rows* and J *columns*. The matrix $\mathbf{A}$ is said to be of *size* $n \times J$. Its generic term is a_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, J$.

When $n = J$, the matrix is said to be *square* of size n. In particular, the basis $v_1, \ldots, v_n$ of F has the representation given by the matrix of size $n \times n$

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & 1 \end{pmatrix}.$$

Such a matrix $\mathbf{I}$ is called the *identity matrix* of size n.

A similar representation can be used for each linear mapping f from E to F. Specifically, the linear mapping f is fully characterized by the knowledge of the images of the basis vectors of E, i.e., by the knowledge of $f(u_1), \ldots, f(u_p)$. The *matrix representation of the linear mapping* f is defined to be the matrix representation of the system $f(u_1), \ldots, f(u_p)$. Hence it is a matrix $\mathbf{A}$ with n rows and p columns whose generic term a_{ij} is the coordinate of the vector $f(u_j)$ associated with the basis vector v_i.

A.1.6 Matrix Operations

When some bases of E and F are given, it is possible to identify any system of vectors of E (or F) or any linear mapping from E to F with a matrix. As a result one can interpret the usual operations on systems and linear mappings in terms of matrix operations.

a) Addition of Matrices

Let f and g be two linear mappings from E to F corresponding to the matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$, respectively. Then $\mathbf{A}$ and $\mathbf{B}$ have the same size. The matrix corresponding to the linear mapping $f + g$ is denoted $\mathbf{C} = \mathbf{A} + \mathbf{B}$. It has the same size as $\mathbf{A}$ and $\mathbf{B}$. Moreover, its generic element is $c_{ij} = a_{ij} + b_{ij}$.

The addition of matrices is commutative and associative. The *zero matrix*, which is the matrix whose elements are all equal to zero, is the neutral element for this operation. Moreover, the addition of two column vectors can be viewed as a special case of the addition of two matrices.

b) Multiplication of a Matrix by a Scalar

Let f be a linear mapping corresponding to the matrix $\mathbf{A} = (a_{ij})$. Let a be a real scalar number. Then af is the linear mapping that corresponds to the matrix $\mathbf{B} = a\mathbf{A}$, which has the same size as $\mathbf{A}$, and whose generic element is $b_{ij} = a \times a_{ij}$. Again the multiplication of a vector by a scalar can be viewed as a special case of this operation.

c) Multiplication of Matrices

Let E, F, and G be three vector spaces of dimension n, p, and q, respectively. Let f be a linear mapping from E to F whose corresponding matrix of size $p \times n$ is $\mathbf{A} = (a_{ij})$. Let g be a linear mapping from F

to G, whose corresponding matrix of size $q \times p$ is $\mathbf{B} = (b_{ij})$. Then the mapping $h = g \circ f$ is linear and corresponds to the matrix of size $q \times n$, denoted $\mathbf{C} = \mathbf{BA}$, whose generic element is $c_{ij} = \sum_{k=1}^{p} b_{ik} a_{kj}$.

Such a multiplication or product is meaningful only if the number of columns of the first matrix $\mathbf{B}$ is equal to the number of rows of the second matrix $\mathbf{A}$. If $p = q = n$, both products $\mathbf{AB}$ and $\mathbf{BA}$ can be computed, though $\mathbf{AB} \neq \mathbf{BA}$, in general.

The product of matrices is also useful for writing the equality $y = f(x)$, where f corresponds to the matrix A. Namely, the matrix formulation of this equality is $y = \mathbf{A}x$.

d) Inverse of a Square Matrix

When E and F have the same dimension, then one may want to know whether the linear mapping f can be inverted. In matrix terms, this is equivalent to answering the following question. Let $\mathbf{A}$ be a square matrix of size n corresponding to f. Does there exist a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix of size n?

It is in fact easy to verify whether a (square) matrix can be inverted. Specifically, it suffices to compute a particular function of the elements of the matrix, called its *determinant*. If $\mathbf{A} = (a_{ij})$ is a square matrix of size n, then its determinant is defined as

$$\det \mathbf{A} = \sum_{\sigma \in \mathcal{S}} (-1)^{\epsilon(\sigma)} \prod_{i=1}^{n} a_{i,\sigma(i)},$$

where $\mathcal{S}$ denotes the set of all permutations of $\{1 \ldots n\}$, and $\epsilon(\sigma)$ denotes the *signature* of the permutation σ.

Then a square matrix $\mathbf{A}$ is *invertible* or *nonsingular* if its determinant is nonzero. When the matrix $\mathbf{A}$ is nonsingular, then its *inverse* exists and is unique. It is denoted $\mathbf{A}^{-1}$.

An important property of the determinant function is

$$\det AB = \det A \det B.$$

Hence the product of two nonsingular matrices is a nonsingular matrix. Moreover, we have $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

e) Transposition of a Matrix

Given a matrix $\mathbf{A} = (a_{ij})$ of size $p \times n$, the *transpose* of $\mathbf{A}$ is the matrix $\mathbf{B} = (b_{ij})$ of size $n \times p$ whose generic element b_{ij} is equal to a_{ji}. The matrix $\mathbf{B}$ is denoted $\mathbf{A}'$.

The transposition operation satisfies the following properties: (i) $(\mathbf{A}+\mathbf{B})' = \mathbf{A}'+\mathbf{B}'$, (ii) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, (iii) $\det \mathbf{A}' = \det \mathbf{A}$ provided $\mathbf{A}$ is a square matrix, and (iv) if $\mathbf{A}$ is nonsingular, then $\mathbf{A}'$ is nonsingular, and $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

A matrix is said to be *symmetric* if $\mathbf{A} = \mathbf{A}'$.

f) Rank of a Matrix

The rank of a matrix is defined to be the rank of the corresponding linear mapping. Then a linear mapping is injective if and only if the rank of its corresponding matrix is equal to the number of its columns, i.e., if the matrix is of *full-column rank.* Similarly a system of vectors is composed of linearly independent vectors if and only if its corresponding matrix is of full-column rank. Lastly, the rank of a matrix is equal to the rank of its transpose, i.e., rank $\mathbf{A}$ = rank $\mathbf{A}'$.

A.1.7 Change of Bases

Given some bases of E and F, we have seen that a unique matrix $\mathbf{A}$ can be associated to any given linear mapping f. When new bases are chosen, however, another matrix corresponds to the mapping f.

a) Change-of-Basis Matrix

Let $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$ be the initial bases of E and F, respectively. Let $(\tilde{u}_1, \ldots, \tilde{u}_p)$ and $(\tilde{v}_1, \ldots, \tilde{v}_n)$ be the new bases of E and F, respectively. For the vector space E, let $\mathbf{P}$ be the $p \times p$ matrix corresponding to the new basis with respect to the old basis of E. Similarly, let $\mathbf{Q}$ be the $n \times n$ matrix corresponding to the new basis of F with respect to the old basis of F. Then both matrices are nonsingular.

b) A New Matrix Representation of f

The matrix corresponding to f with respect to the new bases of E and F is given by $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{AP}$, where $\mathbf{A}$ is the matrix corresponding to f with respect to the old bases. In the special case where $E = F$, $u_j = v_j$, and $\tilde{u}_j = \tilde{v}_j$, the change of bases leads to the matrix representation $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{AQ}$.

c) Diagonalization and Triangularization

In what follows it is assumed that $E = F$. Also, our arguments are derived using the field of complex numbers. In view of the above result we can ask whether there exist new bases for which the matrix representation of f is especially simple. To construct such bases, the concept of eigenvectors is introduced.

An *eigenvector* x of f is a nonzero vector such that $f(x) = \lambda x$. Then λ is called an *eigenvalue* of f. Given λ, the set of eigenvectors is a subspace of E. This set is called the *eigenspace* associated with λ. It is denoted E_λ and is equal to $\ker(f - \lambda \mathbf{I})$. It is non-degenerated in the sense that it is not reduced to $\{0\}$. Thus $\dim \ker(f - \lambda \mathbf{I}) \neq 0$, i.e., $f - \lambda \mathbf{I}$ is singular. This observation suggests to determine the eigenvalues λ as solutions to the so-called *characteristic equation* defined as $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$, where $\mathbf{A}$ is the matrix representing f.

In the sequel the distinct roots of the characteristic equation associated with the matrix $\mathbf{A}$ are denoted $\lambda_1, \ldots, \lambda_K$. We let $\alpha_1, \ldots, \alpha_K$ denote the multiplicity of these roots. We have $\alpha_1 + \ldots + \alpha_K = n$. Two cases must be distinguished:

(i) If $\alpha_k = \dim E_{\lambda_k}$, $\forall k = 1, \ldots, K$, then it is possible to construct a basis that consists of eigenvectors only. With respect to such a basis, the linear mapping f can be represented by the diagonal matrix

$$\mathbf{\Delta} = \begin{pmatrix} \lambda_1 & & & & 0 \\ & \ddots & & & \\ & & \lambda_1 & & \\ & & & \ddots & \\ & & & & \lambda_K \\ & 0 & & & & \ddots \\ & & & & & & \lambda_K \end{pmatrix} \begin{matrix} \} \alpha_1 \\ \\ \vdots \\ \\ \\ \} \alpha_K \end{matrix}$$

Thus, with respect to any other basis, the matrix representing f can be written as $\mathbf{A} = \mathbf{Q}^{-1} \mathbf{\Delta} \mathbf{Q}$.

(ii) If there exists a k such that $\dim E_{\lambda_k} < \alpha_k$, then it is not possible to find a basis that consists of eigenvectors only. Nonetheless it is possible to construct a basis for which the matrix representing f is lower

triangular with eigenvalues on its main diagonal, i.e

$$\mathbf{T} = \begin{pmatrix} \lambda_1 & & & & & 0 & \\ & \ddots & & & & & \\ & & \lambda_1 & & & & \\ & & & \ddots & & & \\ & & & & \lambda_K & & \\ & & & & & \ddots & \\ & & & & & & \lambda_K \end{pmatrix} \begin{matrix} \} \alpha_1 \\ \\ \vdots \\ \\ \} \alpha_K \end{matrix}$$

Moreover, any other matrix representation of the linear mapping f is of the form $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}$. It follows that

$$\det(\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}) = \det \mathbf{T} = \prod_{i=1}^{n} \lambda_i = \prod_{k=1}^{K} \lambda_k^{\alpha_k}.$$

A.2 Some Complements on Matrices

A.2.1 Trace of a Matrix

Definition A.1: *Given a square matrix* $\mathbf{A}$ *of size* n *with generic element* a_{ij}*, the trace of* $\mathbf{A}$ *denoted* tr $\mathbf{A}$ *is the sum of its diagonal elements, i.e.*

$$\text{tr } \mathbf{A} = \sum_{i=1}^{n} a_{ii}.$$

It is easy to verify that the trace operator is linear, i.e.

$$\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr } \mathbf{A} + \text{tr } \mathbf{B}, \ \text{tr}(\lambda \mathbf{A}) = \lambda \text{tr } \mathbf{A}.$$

Moreover

$$\text{tr } \mathbf{A}' = \text{tr } \mathbf{A}.$$

Property A.1: *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two matrices of size* $n \times p$ *and* $p \times n$*, respectively. Then the two matrices* $\mathbf{AB}$ *and* $\mathbf{BA}$ *are square matrices of size* $n \times n$ *and* $p \times p$*, respectively. Moreover, these matrices have the same trace, i.e.,* $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.

PROOF: Let $\mathbf{C} = \mathbf{AB}$ and $\mathbf{D} = \mathbf{BA}$. We have

$$\text{tr } \mathbf{C} = \sum_{i=1}^{n} c_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij} b_{ji} = \sum_{j=1}^{p} \sum_{i=1}^{n} b_{ji} a_{ij} = \sum_{j=1}^{p} d_{jj} = \text{tr } \mathbf{D}.$$

□

Property A.2: *The trace of a matrix is invariant with respect to a change of basis, i.e.*

$$\text{tr}(\mathbf{Q}^{-1}\mathbf{AQ}) = \text{tr } \mathbf{A}.$$

PROOF: From Property A.1, we have

$$\text{tr}(\mathbf{Q}^{-1}\mathbf{AQ}) = \text{tr}(\mathbf{AQQ}^{-1}) = \text{tr}(\mathbf{AI}) = \text{tr } \mathbf{A}.$$

□

Property A.3: *The trace of a matrix is equal to the sum of its eigenvalues (counting multiplicities).*

PROOF: As seen earlier, there always exists a triangular matrix representation of the linear mapping associated with the given matrix $\mathbf{A}$. Hence, from Property A.2, we obtain

$$\text{tr } \mathbf{A} = \text{tr } \mathbf{T} = \alpha_1 \lambda_1 + \ldots + \alpha_K \lambda_K.$$

□

A.2.2 Partitioned Matrices

A matrix $\mathbf{A}$ is frequently written by displaying its elements a_{ij} in a rectangular array. For instance, we have

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 & 0 & 3 \\ 2 & 1 & 1 & 0 & 1 \\ 3 & 1 & 2 & 1 & 0 \end{pmatrix}.$$

There are many other ways to display the matrix $\mathbf{A}$. In particular, the matrix $\mathbf{A}$ can be displayed by defining the *submatrices* or *blocks* that compose it. For instance, we can have

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

with

$$\mathbf{A}_{11} = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 1 \end{pmatrix}, \quad \mathbf{A}_{12} = \begin{pmatrix} 0 & 3 \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{A}_{21} = (3, 1, 2), \qquad \mathbf{A}_{22} = (1, 0).$$

More generally, a *partitioning* of $\mathbf{A}$ is

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \dots & \mathbf{A}_{1L} \\ \vdots & & \vdots \\ \mathbf{A}_{K1} & \dots & \mathbf{A}_{KL} \end{pmatrix},$$

where the submatrice $\mathbf{A}_{kl}$ is of size $n_k \times p_l$, and $n = \sum_{k=1}^{K} n_k$, $p = \sum_{l=1}^{L} p_l$. Note that all blocks in the same column have the same number of columns. Similarly, all blocks in the same row have the same number of rows.

Among all the possible partitionings of a matrix, some are more useful than others. In particular, we shall frequently use partitionings of the form $\mathbf{A} = (A_1, \dots, A_p)$, where the A_j's are vectors of size n. These vectors are called the *column vectors* of $\mathbf{A}$. They are simply the images of the canonical basis vectors by the linear mapping associated with $\mathbf{A}$.

A partitioned display has the advantage of being concise. Within such a context, it is useful to extend the usual matrix operations such as the transposition, the addition, the multiplication, and the inversion. These operations are discussed next when matrices are partitioned in four blocks such as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad n_1 + n_2 = n, \quad p_1 + p_2 = p.$$

The generalization to arbitrary partitionings is straightforward.

a) Transposition

We have

$$\mathbf{A}' = \begin{pmatrix} \mathbf{A}'_{11} & \mathbf{A}'_{21} \\ \mathbf{A}'_{12} & \mathbf{A}'_{22} \end{pmatrix},$$

where $\mathbf{A}'$ is of size $p \times n$. Note that it is necessary to transpose the partitioned matrix as well as each of its blocks.

In particular, a symmetric matrix $\mathbf{\Omega}$ of size n can be written as

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}'_{12} & \mathbf{\Omega}_{22} \end{pmatrix},$$

where $\mathbf{\Omega}_{11}$ and $\mathbf{\Omega}_{22}$ are two symmetric matrices.

b) Addition

We have

$$\begin{aligned}\mathbf{A}+\mathbf{B} &= \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} + \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}_{11}+\mathbf{B}_{11} & \mathbf{A}_{12}+\mathbf{B}_{12} \\ \mathbf{A}_{21}+\mathbf{B}_{21} & \mathbf{A}_{22}+\mathbf{B}_{22} \end{pmatrix},\end{aligned}$$

where $\mathbf{A}$ and $\mathbf{B}$ are of size $n \times p$. Of course, the submatrices of $\mathbf{A}$ and $\mathbf{B}$ corresponding to the same subscripts must have the same size for the additions to have a sense.

c) Multiplication by Blocks

Let

$$\mathbf{C} = \mathbf{AB},$$

where $\mathbf{C}$, $\mathbf{A}$, and $\mathbf{B}$ are of size $n \times q$, $n \times p$, and $p \times q$, respectively. Let

$$\mathbf{A} = \begin{pmatrix} \underset{n_1 \times p_1}{\mathbf{A}_{11}} & \underset{n_1 \times p_2}{\mathbf{A}_{12}} \\ \\ \underset{n_2 \times p_1}{\mathbf{A}_{21}} & \underset{n_2 \times p_2}{\mathbf{A}_{22}} \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} \underset{p_1 \times q_1}{\mathbf{B}_{11}} & \underset{p_1 \times q_2}{\mathbf{B}_{12}} \\ \\ \underset{p_2 \times q_1}{\mathbf{B}_{21}} & \underset{p_2 \times q_2}{\mathbf{B}_{22}} \end{pmatrix},$$

where $n_1 + n_2 = n$, $p_1 + p_2 = p$, and $q_1 + q_2 = q$. Then

$$\mathbf{C} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11}+\mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12}+\mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11}+\mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12}+\mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix}.$$

Note that the formula giving the product of partitioned matrices is formally identical to the one that would be obtained if the blocks $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ had only one element a_{ij} and b_{ij} each. Note, however, that the product of submatrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ should not be reversed. In addition, the sizes of the various submatrices should be compatible for the various matrix multiplications to have a sense.

The preceding product of partitioned matrices is for instance useful for specifying that a vector belongs to a subspace generated by some given vectors $X_1, \ldots, X_p \in \mathbb{R}^n$. Specifically, any vector Y of such a subspace can be written as $Y = X_1 b_1 + \ldots + X_p b_p$, where $b_1, \ldots, b_p$ are

the corresponding coordinates of Y. Then, letting

$$\mathbf{X} = (X_1 \dots X_p), \text{ and } b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix},$$

we see that $Y = \mathbf{X}b$.

d) Inverse of a Partitioned Matrix

Let $\mathbf{A}$ be a square matrix of size n whose two diagonal blocks are square submatrices, namely

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are respectively of size $n_1 \times n_1$, $n_1 \times n_2$, $n_2 \times n_1$, and $n_2 \times n_2$ with $n_1 + n_2 = n$.

If $\mathbf{A}$ is nonsingular, then its inverse matrix $\mathbf{A}^{-1}$ can be partitioned as

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix},$$

where $\mathbf{A}^{11}$, $\mathbf{A}^{12}$, $\mathbf{A}^{21}$, and $\mathbf{A}^{22}$ are respectively of size $n_1 \times n_1$, $n_1 \times n_2$, $n_2 \times n_1$, and $n_2 \times n_2$ with $n_1 + n_2 = n$. To determine $\mathbf{A}^{-1}$, it is necessary to express the blocks $\mathbf{A}^{kl}$ of the inverse matrix in terms of the blocks $\mathbf{A}_{kl}$ of the original matrix.

To do so, it suffices to use the matrix equality $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ and to partition the identity matrix $\mathbf{I}$ accordingly. Then, using the formula for the multiplication by blocks, we obtain four equations in the four unknown matrices $\mathbf{A}^{11}$, $\mathbf{A}^{12}$, $\mathbf{A}^{21}$, and $\mathbf{A}^{22}$

$$\begin{aligned} \mathbf{A}_{11}\mathbf{A}^{11} + \mathbf{A}_{12}\mathbf{A}^{21} &= \mathbf{I}, \\ \mathbf{A}_{11}\mathbf{A}^{12} + \mathbf{A}_{12}\mathbf{A}^{22} &= 0, \\ \mathbf{A}_{21}\mathbf{A}^{11} + \mathbf{A}_{22}\mathbf{A}^{21} &= 0, \\ \mathbf{A}_{21}\mathbf{A}^{12} + \mathbf{A}_{22}\mathbf{A}^{22} &= \mathbf{I}. \end{aligned}$$

Assuming that $\mathbf{A}_{11}$ is *nonsingular*, we obtain from the first equation

$$\mathbf{A}^{11} = \mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{21}.$$

Substituting this expression in the third equation, we obtain

$$\mathbf{A}_{21}\left(\mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{21}\right) + \mathbf{A}_{22}\mathbf{A}^{21} = 0,$$

i.e.

$$\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)\mathbf{A}^{21} = -\mathbf{A}_{21}\mathbf{A}_{11}^{-1}.$$

Moreover, if $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ is *nonsingular*, it follows that

$$\begin{aligned}
\mathbf{A}^{21} &= -\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}, \\
\mathbf{A}^{11} &= \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}.
\end{aligned}$$

Computation of the remaining blocks relies upon the remaining two equalities of the above system. For instance, from the second equality we obtain: $\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22}$. Then, substituting this expression in the fourth equation, we obtain

$$(-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} + \mathbf{A}_{22})\mathbf{A}^{22} = \mathbf{I}.$$

Hence

$$\mathbf{A}^{22} = \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}.$$

Thus

$$\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}.$$

To summarize the preceding results, we have

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{E}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{E} \\ -\mathbf{E}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{E} \end{pmatrix},$$

where

$$\mathbf{E} = \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}.$$

e) A Useful Property

Property A.4: *We have*

$$\begin{aligned}
&\mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} \\
&\quad = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1}
\end{aligned}$$

PROOF: Both sides of the preceding equation are equal to $\mathbf{A}^{11}$. The expression for $\mathbf{A}^{11}$ in the left-hand side was obtained previously. The expression in the right-hand side follows from the expression for $\mathbf{A}^{22}$ by reversing the role of the subscripts 1 and 2. □

f) Determinant of a Partitioned Matrix

We consider again a square matrix $\mathbf{A}$ whose two diagonal blocks $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square matrices.

Property A.5:

(i) If $\mathbf{A}_{11}$ is nonsingular, then

$$\det \mathbf{A} = \det \mathbf{A}_{11} \det \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right).$$

(ii) If $\mathbf{A}_{22}$ is nonsingular, then

$$\det \mathbf{A} = \det \mathbf{A}_{22} \det \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right).$$

PROOF: We only prove (i), (ii) being proved similarly. To prove (i), it suffices to multiply the matrix $\mathbf{A}$ by an appropriately chosen matrix so that the resulting product is block diagonal. Specifically, we have

$$\begin{pmatrix} \mathbf{I} & 0 \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \end{pmatrix} \mathbf{A} \begin{pmatrix} \mathbf{I} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ 0 & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & 0 \\ 0 & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{pmatrix}.$$

Because the determinant of a block diagonal matrix is equal to the product of the determinants of the diagonal blocks, and because the determinant of a triangular matrix is equal to the product of the diagonal elements, then we have

$$\det \mathbf{A} = \det \mathbf{A}_{11} \det \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right).$$

□

In particular, Property A.5 implies that it was unnecessary to assume that the matrices $\mathbf{A}_{11}$ and $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ are *both* nonsingular, as it was assumed when deriving the inverse $\mathbf{A}^{-1}$ of the partitioned matrix $\mathbf{A}$. For, if $\mathbf{A}$ and $\mathbf{A}_{11}$ are nonsingular, then $\det \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right) = (\det \mathbf{A}/\det \mathbf{A}_{11}) \neq 0$. Hence $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ is necessarily nonsingular.

Note also that $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ may be both singular. For instance, take

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In this case the preceding formulae for computing the partitioned inverse matrix $\mathbf{A}^{-1}$ and the determinant of the matrix $\mathbf{A}$ no longer apply.

Lastly, note that the blocks of the partitioned inverse matrix can be similarly computed from $\mathbf{A}_{22}^{-1}$, when $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are both nonsingular. In this case, we obtain alternative expressions for these blocks, which are necessarily equal to those found earlier. For instance, we have

$$\begin{aligned}&\left(\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}\\ &\quad=\ \mathbf{A}_{22}^{-1}+\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\left(\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}. \quad \text{(A.1)}\end{aligned}$$

A.3 A Review of Quadratic Forms

A.3.1 Definition and Characterization

Let E be a vector space of dimension n. A *symmetric bilinear form* ω is a mapping from $E \times E$ to $\mathbb{R}$ satisfying:

(i) The symmetry condition

$$\forall x, y \in E : \omega(x, y) = \omega(y, x),$$

(ii) The linearity in each of its arguments

$$\begin{aligned}\forall\ x, y, z \in E,\ \forall\ a, b \in \mathbb{R}: \qquad &\omega(ax + bz, y) = a\omega(x, y) + b\omega(z, y),\\ &\omega(x, ay + bz) = a\omega(x, y) + b\omega(x, z).\end{aligned}$$

To each symmetric bilinear form one can associate the mapping $\mathbf{\Omega}$ from E to $\mathbb{R}$ defined by $\mathbf{\Omega}(\mathbf{x}) = \omega(x, x), \forall x \in E$. Then $\mathbf{\Omega}$ is called the *quadratic form* associated with ω. As a matter of fact, it is equivalent to know either one of these two mappings.

Specifically, when the space E is endowed with a basis $u_1, \ldots, u_n$, the bilinear form is fully characterized by the values $\omega_{ij} = \omega(u_i, u_j)$, $i, j = 1, \ldots, n$, taken by this form when evaluated at the basis vectors. Let $\mathbf{\Omega}$ be the square matrix of size n whose elements are the ω_{ij}'s. It is easy to see that $\mathbf{\Omega}$ is a symmetric matrix. Moreover, for every pair of vectors x and y with matrix representations x and y, we have

$$\omega(x, y) = x'\mathbf{\Omega}y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j \omega_{ij}$$

and

$$\Omega(x) = x'\Omega x = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \omega_{ij}.$$

Conversely, every symmetric matrix Ω defines a bilinear form through the formula $\omega(x, y) = x'\Omega y$.

When the basis changes, then the matrix associated with the given symmetric bilinear form also changes. It becomes $\mathbf{Q}'\Omega\mathbf{Q}$, where $\mathbf{Q}$ is the matrix of the change of basis.

A.3.2 Positive Quadratic Forms

A quadratic form Ω is said to be *positive semidefinite* if and only if $\Omega(x) \geq 0, \forall x \in E$. It is said to be *positive definite* if, in addition, $\Omega(x) = 0 \Rightarrow x = 0$.

When the quadratic form is positive semidefinite, then the following inequality holds

$$\forall\, x, y \in E : \omega(x, y)^2 \leq \Omega(x)\Omega(y),$$

which is called the *Schwarz inequality*. In particular, it implies that, if x satisfies $\Omega(x) = 0$, then $\omega(x, y) = 0$, $\forall\, y$. Hence $\Omega x = 0$, i.e., $x \in \ker \Omega$. Conversely, every element of $\ker \Omega$ clearly satisfies $\Omega(x) = 0$.

The Schwarz inequality becomes an equality, i.e., $\omega(x, y)^2 = \Omega(x)\,\Omega(y)$, if and only if there exists a $\lambda \in I\!R$: $y - \lambda x \in \ker \Omega$.

When Ω is positive definite, it is possible to define some concepts of orthogonality and distance. Specifically, the mapping defined by $\|x\| = \sqrt{\Omega(x)}$ satisfies all the properties of a norm. Then the corresponding distance is defined by $d(x, y) = \|x - y\|$. This norm is called a *Euclidean norm*. In this case it is customary to use a simpler notation by replacing $\Omega(x)$ by $\|x\|^2$ and $\omega(x, y)$ by $< x, y >$, which is called the *scalar product* associated with the form ω.

In particular, two vectors x and y are said to be orthogonal to each other if $< x, y >= 0$. Such an orthogonality property is commonly represented by a right angle, as displayed in Figure A.1. The concept of orthogonality depends, however, upon the quadratic form used. Thus two vectors that are orthogonal to each other with respect to a given form Ω are in general not orthogonal with respect to another quadratic form Ω'.

A well-known characterization of orthogonality is provided by the next result.

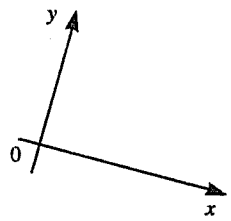


Figure A.1: Two Orthogonal Vectors

Theorem A.1: Pythagora's Theorem. *We have*

$$< x, y > = 0 \Leftrightarrow \|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

The concept of orthogonality of two vectors can be generalized to two subspaces. Let F and G be two subspaces of E. Then F and G are *orthogonal* if $\forall\ x \in F$, $\forall\ y \in G$: $< x, y >= 0$. If F is a subspace of E, then the space of vectors that are orthogonal to all vectors of F is a subspace. It is called the *subspace orthogonal to* F and is denoted $F^{\perp}$. An example is illustrated in Figure A.2. Moreover, we have $E = F \oplus F^{\perp}$. Hence $F \cap F^{\perp} = \{0\}$ and $\dim F + \dim F^{\perp} = n$.

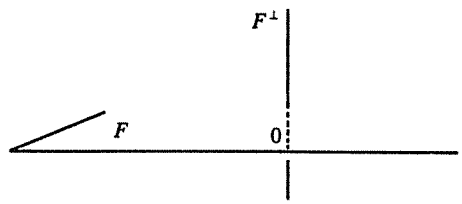


Figure A.2: Orthogonal Subspaces

The concept of orthogonality can also be generalized to more than two elements. A *system* of vectors or subspaces is said to be orthogonal if any pair of vectors or subspaces are mutually orthogonal. Note that any system of orthogonal nonzero vectors is linearly independent. Thus it is natural to ask whether such as system can be augmented so as to constitute a basis. The answer is positive. Specifically, every vector space endowed with a scalar product contains a basis of orthogonal vectors. In fact, there is an infinity of such orthogonal bases. The basis vectors

of any such basis can *normalized* to have a length equal to one. The resulting basis is said to be *orthonormal.*

Any matrix $\mathbf{Q}$ corresponding to a change from an orthonormal basis to another orthonormal basis has the following property: $\mathbf{Q}^{-1} = \mathbf{Q}'$. Such matrices are said to be *orthogonal.*

A.3.3 Symmetric Matrix Representations of a Linear Mapping

A symmetric matrix $\mathbf{\Omega}$ can be viewed as representing a quadratic form. It can also be viewed as representing a linear mapping. In the latter case the matrix $\mathbf{\Omega}$ is invariant with respect to transposition. Hence, with such an interpretation, we can consider the problem of diagonalizing $\mathbf{\Omega}$.

Such a symmetric matrix can be diagonalized with eigenvalues that are all real. Moreover, the eigenvectors can be chosen so as to constitute an orthonormal basis. Thus we have $\mathbf{\Omega} = \mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{Q} = \mathbf{Q}'\mathbf{\Delta}\mathbf{Q}$ since $\mathbf{Q}$ is an orthogonal matrix. With respect to this orthonormal basis the matrix $\mathbf{\Delta}$ represents simultaneously the linear mapping and the quadratic form associated with $\mathbf{\Omega}$.

In particular, the matrix $\mathbf{\Omega}$ corresponds to a positive semidefinite quadratic form if and only if $\mathbf{\Delta}$ is positive semidefinite, i.e., if and only all the eigenvalues are nonnegative. It is positive definite if all the eigenvalues are strictly positive. It follows that the determinant of such a matrix is strictly positive.

A.3.4 Relationships among $\ker \mathbf{A}$, $\mathrm{Im}\ \mathbf{A}$, $\ker \mathbf{A}'$, and $\mathrm{Im}\ \mathbf{A}'$

Let E be a vector space endowed with a basis and scalar product $< \cdot, \cdot >$ whose matrix representation with respect to such a basis is the identity matrix. We have

$$\forall\, x, y \in E : < x, y > = x'y.$$

Let f and f' be two elements of the space of linear mappings $\mathcal{L}(E, E)$, whose corresponding matrix representations are $\mathbf{A}$ and $\mathbf{A}'$. We have

$$< f(x), y > = < x, f'(y) > = x'\mathbf{A}'y.$$

Property A.6: *We have*

(i) $\ker \mathbf{A}' = (\text{Im } \mathbf{A})^{\perp}$,

(ii) $\text{Im } \mathbf{A}' = (\ker \mathbf{A})^{\perp}$.

PROOF: First, let $x \in \ker \mathbf{A}$ and $y \in \text{Im } \mathbf{A}'$. Then there exists z such that $y = f'(z)$. Hence we have $< x, y > = < x, f'(z) > = < f(x), z > = 0$ because $f(x) = 0$ by assumption. It follows that the two subspaces $\ker \mathbf{A}$ and $\text{Im } \mathbf{A}'$ are mutually orthogonal. Hence $\text{Im } \mathbf{A}' \subset (\ker \mathbf{A})^{\perp}$.

On the other hand, we know that $\text{rank } \mathbf{A}' = \text{rank } \mathbf{A} = \dim E - \dim \ker \mathbf{A}$. Or equivalently, we know that $\dim \text{Im } \mathbf{A}' = \dim(\ker \mathbf{A})^{\perp}$. It follows that $\text{Im } \mathbf{A}' = (\ker \mathbf{A})^{\perp}$.

Lastly, (ii) follows from $\text{Im } \mathbf{A} = (\ker \mathbf{A}')^{\perp}$ by considering their orthogonal spaces. □

A.4 Ordering Symmetric Matrices

The concept of positive semidefiniteness for symmetric matrices can be used to define a (partial) order relation on the set of symmetric matrices. Throughout this section we assume that all matrices $\mathbf{\Omega}$, $\mathbf{\Omega}_1$, $\mathbf{\Omega}_2$, ... are symmetric matrices of same size n. We denote by $\mathbf{\Omega} \succeq 0$ the fact that the matrix $\mathbf{\Omega}$ corresponds to a positive semidefinite quadratic form.

A.4.1 Definition of an Order Relation

Definition A.2: *A symmetric matrix $\mathbf{\Omega}_1$ is said to be larger than another symmetric matrix $\mathbf{\Omega}_2$ if their difference $\mathbf{\Omega}_1 - \mathbf{\Omega}_2$ is positive semidefinite. Formally, this is written*

$$\begin{aligned} \mathbf{\Omega}_1 \succeq \mathbf{\Omega}_2 &\Leftrightarrow \mathbf{\Omega}_1 - \mathbf{\Omega}_2 \succeq 0 \\ &\Leftrightarrow \forall\, x \in E, x'\mathbf{\Omega}_1 x \geq x'\mathbf{\Omega}_2 x. \end{aligned}$$

The relation $\succeq$ is analogous to the usual order relation $\geq$ on $\mathbb{R}$. It has the following properties:

(i) $\mathbf{\Omega} \succeq \mathbf{\Omega}$ (reflexive),

(ii) $\mathbf{\Omega}_1 \succeq \mathbf{\Omega}_2$ and $\mathbf{\Omega}_2 \succeq \mathbf{\Omega}_1 \Rightarrow \mathbf{\Omega}_1 = \mathbf{\Omega}_2$ (antisymmetric),

(iii) $\mathbf{\Omega}_1 \succeq \mathbf{\Omega}_2$ and $\mathbf{\Omega}_2 \succeq \mathbf{\Omega}_3 \Rightarrow \mathbf{\Omega}_1 \succeq \mathbf{\Omega}_3$ (transitive),

(iv) The order relation $\succeq$ is not total.

(v) $\Omega_1 \succeq \Omega_2 \Leftrightarrow -\Omega_2 \succeq -\Omega_1$,

(vi) $\forall\, \lambda \in I\!R^{+*}$, $\Omega_1 \succeq \Omega_2 \Leftrightarrow \lambda\Omega_1 \succeq \lambda\Omega_2$,

(vii) $\forall\, \Omega$ symmetric, $\Omega_1 \succeq \Omega_2 \Leftrightarrow \Omega + \Omega_1 \succeq \Omega + \Omega_2$,

(viii) $\Omega_1 \succeq \Omega_2 \Rightarrow \mathbf{Q}\Omega_1\mathbf{Q}' \succeq \mathbf{Q}\Omega_2\mathbf{Q}'$, $\forall\, \mathbf{Q}$.

A.4.2 Comparing Ordered Spectra

The *ordered spectrum* of a symmetric matrix Ω (say) is defined as the set of its eigenvalues ranked in decreasing order, i.e., $\text{Sp}\ \Omega = (\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n)$.

Definition A.3: *Consider two symmetric matrices Ω_1 and Ω_2 with respective ordered spectra*

$$\begin{aligned} \text{Sp}\ \Omega_1 &= (\lambda_{11} \geq \lambda_{12} \ldots \geq \lambda_{1n}), \\ \text{Sp}\ \Omega_2 &= (\lambda_{21} \geq \lambda_{22} \ldots \geq \lambda_{2n}). \end{aligned}$$

The spectrum of the first matrix is at least as large as the spectrum of the second matrix, which is denoted $\text{Sp}\ \Omega_1 \geq \text{Sp}\ \Omega_2$, *if* $\lambda_{1j} \geq \lambda_{2j}$, $\forall\, j = 1, \ldots, n$.

To compare this order relation to the preceding order relation $\succeq$, the next result can be useful. It is also interesting by itself.

Property A.7: *Let Ω be a symmetric matrix. Let $e_1, \ldots, e_p$ be some orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$. Let F_p denote the subspace generated by $e_1, \ldots, e_p$.*

We have

$$\begin{aligned} \max_{x \in F_p} \frac{x'\Omega x}{x'x} &= \lambda_1, \\ \min_{x \in F_p} \frac{x'\Omega x}{x'x} &= \lambda_p. \end{aligned}$$

PROOF: Every vector $x \in F$ can be written as $x = \sum_{j=1}^{p} x_j e_j$. Because the e_j's are eigenvectors that constitute an orthonormal system, then we have

$$\frac{x'\Omega x}{x'x} = \frac{\sum_{j=1}^{p} x_j^2 \lambda_j}{\sum_{j=1}^{p} x_j^2}.$$

The desired result follows at once. □

An immediate consequence is as follows.

Corollary A.1: *We have*

$$\max_{x \in E} \frac{x'\Omega x}{x'x} = \max_i \lambda_i, \ \min_{x \in E} \frac{x'\Omega x}{x'x} = \min_i \lambda_i.$$

Property A.8: *We have*

$$\Omega_1 \succeq \Omega_2 \Rightarrow \text{Sp } \Omega_1 \geq \text{Sp } \Omega_2.$$

PROOF: Let F_1 be the subspace generated by the eigenvectors associated with the $n-p+1$ smallest eigenvalues of Ω_1. Let F_2 be the subspace generated by the eigenvectors associated with the p largest eigenvalues of Ω_2. From Property A.7 we have

$$\begin{aligned} \max_{x \in F_1} \frac{x'\Omega_1 x}{x'x} &= \lambda_{1p}, \\ \min_{x \in F_2} \frac{x'\Omega_2 x}{x'x} &= \lambda_{2p}. \end{aligned}$$

Because $\dim F_1 + \dim F_2 = n+1 > n$, then the two subspaces F_1 and F_2 have an intersection that is not reduced to $\{0\}$. Let x_0 be a nonzero element of this intersection. We have $x_0'\Omega_1 x_0 \geq x_0'\Omega_2 x_0$. Hence

$$\lambda_{1p} = \max_{x \in F_1} \frac{x'\Omega_1 x}{x'x} \geq \frac{x_0'\Omega_1 x_0}{x_0'x_0} \geq \frac{x_0'\Omega_2 x_0}{x_0'x_0} \geq \min_{x \in F_2} \frac{x'\Omega_2 x}{x'x} = \lambda_{2p}.$$

□

Corollary A.2: *If* $\Omega_1 \succeq \Omega_2 \succeq 0$, *then* $\det \Omega_1 \geq \det \Omega_2$.

PROOF: Because $\forall\, i, \lambda_{1i} \geq \lambda_{2i} \geq 0$, we have

$$\det \Omega_1 = \prod_{i=1}^{n} \lambda_{1i} \geq \prod_{i=1}^{n} \lambda_{2i} = \det \Omega_2.$$

□

Property A.9: *We have* $\text{Sp } \Omega_1 \geq \text{Sp } \Omega_2$ *if and only if there exists an orthogonal matrix* $\mathbf{Q}$ *such that* $\Omega_1 \succeq \mathbf{Q}'\Omega_2\mathbf{Q}$.

PROOF: First, because the properties of the order relation $\succeq$ do not depend upon the chosen basis in view of property (viii), we can choose to represent $\mathbf{\Omega}_1$ by the diagonal matrix

$$\begin{pmatrix} \lambda_{11} & & 0 \\ & \ddots & \\ 0 & & \lambda_{1n} \end{pmatrix}.$$

To prove necessity, it suffices to note that the inequality Sp $\mathbf{\Omega}_1 \geq$ Sp $\mathbf{\Omega}_2$ implies that the above diagonal form is larger (in the sense of the order relation $\succeq$) than the diagonal matrix representing $\mathbf{\Omega}_2$, which is

$$\begin{pmatrix} \lambda_{21} & & 0 \\ & \ddots & \\ 0 & & \lambda_{2n} \end{pmatrix}.$$

The desired conclusion is obtained by letting $\mathbf{Q}$ be the matrix representing an orthonormal basis of eigenvectors of $\mathbf{\Omega}_2$.

To prove sufficiency, we note that, if $\mathbf{\Omega}_1 \succeq \mathbf{Q}'\mathbf{\Omega}_2\mathbf{Q}$, then

$$\text{Sp } \mathbf{\Omega}_1 \geq \text{Sp}(\mathbf{Q}'\mathbf{\Omega}_2\mathbf{Q}) = \text{Sp } \mathbf{\Omega}_2.$$

□

A.4.3 Square Root of a Positive Semidefinite Matrix

Property A.10: *Let $\mathbf{\Omega}$ be a symmetric positive semidefinite matrix. Then there exists a symmetric positive semidefinite matrix $\mathbf{A}$ such that $\mathbf{\Omega} = \mathbf{A}^2$. Such a matrix is unique and is denoted $\mathbf{\Omega}^{1/2}$. Moreover, it is nonsingular if $\mathbf{\Omega}$ is nonsingular, in which case $(\mathbf{\Omega}^{1/2})^{-1} = (\mathbf{\Omega}^{-1})^{1/2}$. The latter matrix is denoted $\mathbf{\Omega}^{-1/2}$.*

PROOF: (i) We can write $\mathbf{\Omega}$ as

$$\mathbf{\Omega} = \mathbf{Q}' \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \mathbf{Q},$$

where $\mathbf{Q}$ is orthogonal and $\lambda_i \geq 0$, $\forall\ i$. Hence

$$\begin{aligned} \mathbf{\Omega} &= \mathbf{Q}' \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix} \mathbf{Q}\mathbf{Q}' \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix} \mathbf{Q} \\ &= \mathbf{A}^2, \end{aligned}$$

where

$$\mathbf{A} = \mathbf{Q}' \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix} \mathbf{Q}.$$

It is easy to verify that $\mathbf{A}$ is symmetric. It is also positive semidefinite since all its eigenvalues, which are equal to $\sqrt{\lambda_i}$, are all nonnegative.

(ii) The uniqueness of $\mathbf{A}$ follows from the fact that $\mathbf{\Omega} = \mathbf{A}^2$ has necessarily the same eigenvectors as $\mathbf{A}$ with eigenvalues equal to the squares of the eigenvalues of $\mathbf{A}$.

(iii) The nonsingularity of $\mathbf{\Omega}$ is equivalent to

$$\lambda_i > 0, \ \forall \ i,$$

i.e., to

$$\sqrt{\lambda_i} > 0, \ \forall \ i,$$

i.e., to the nonsingularity of $\mathbf{\Omega}^{1/2}$.

(iv) When $\mathbf{\Omega}$ is nonsingular, then the inverse of $\mathbf{\Omega}^{1/2}$ is

$$\begin{aligned} (\mathbf{\Omega}^{1/2})^{-1} &= \mathbf{Q}^{-1} \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\lambda_n}} \end{pmatrix} \mathbf{Q}'^{-1} \\ &= \mathbf{Q}' \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\lambda_n}} \end{pmatrix} \mathbf{Q} = (\mathbf{\Omega}^{-1})^{1/2}. \end{aligned}$$

□

Corollary A.3: *Suppose that* $\mathbf{\Omega}_2 \succeq \mathbf{\Omega}_1 \succeq 0$ *and that* $\mathbf{\Omega}_1$ *is nonsingular. Then* $\mathbf{\Omega}_1^{-1} \succeq \mathbf{\Omega}_2^{-1}$.

PROOF: First, we note that Sp $\mathbf{\Omega}_2 \geq$ Sp $\mathbf{\Omega}_1$ by Property A.8. Hence all the eigenvalues of $\mathbf{\Omega}_2$ are at least as large as the smallest eigenvalue of $\mathbf{\Omega}_1$. Because the latter eigenvalue is strictly positive, then all the eigenvalues of $\mathbf{\Omega}_2$ are strictly positive. Thus $\mathbf{\Omega}_2$ is nonsingular.

Next, the condition $\mathbf{\Omega}_2 \succeq \mathbf{\Omega}_1$ is equivalent to $\mathbf{P}'\mathbf{\Omega}_2\mathbf{P} \succeq \mathbf{P}'\mathbf{\Omega}_1\mathbf{P}$, where $\mathbf{P} = \mathbf{\Omega}_1^{-1/2}$ for instance. Hence $\mathbf{\Omega}_2 \succeq \mathbf{\Omega}_1$ is equivalent to $\mathbf{\Omega} \equiv \mathbf{\Omega}_1^{-1/2}\mathbf{\Omega}_2\mathbf{\Omega}_1^{-1/2} \succeq \mathbf{I}$. Similarly, the conclusion $\mathbf{\Omega}_1^{-1} \succeq \mathbf{\Omega}_2^{-1}$ is equivalent to

$$\mathbf{I} \succeq \mathbf{\Omega}^{-1} = \mathbf{\Omega}_1^{1/2}\mathbf{\Omega}_2^{-1}\mathbf{\Omega}_1^{1/2}.$$

Then the corollary follows from the next lemma.

Lemma A.1: *If* $\Omega \succeq \mathbf{I}$, *then* $\mathbf{I} \succeq \Omega^{-1}$.

PROOF: Using Corollary A.1, we have

$$\begin{aligned} \Omega \succeq \mathbf{I} &\Leftrightarrow \forall\, x \in E, x'\Omega x \geq x'x \\ &\Leftrightarrow \min_{x \in E} \frac{x'\Omega x}{x'x} = \min_i \lambda_i \geq 1. \end{aligned}$$

Similarly

$$\mathbf{I} \succeq \Omega^{-1} \Leftrightarrow \max_{x \in E} \frac{x'\Omega^{-1}x}{x'x} = \max_i \frac{1}{\lambda_i} = \frac{1}{\min_i \lambda_i} \leq 1.$$

The two right-hand side conditions are clearly equivalent. □

A.4.4 A Geometric Interpretation

Let Ω be a symmetric positive definite matrix. Consider the vectors satisfying $x'\Omega x = 1$. Provided we choose a suitable change of orthonormal basis, we can assume that Ω is diagonal. Thus, with respect to this new basis, the preceding equality becomes $\sum_{i=1}^n \lambda_i \tilde{x}_i^2 = 1$, where $\tilde{x}_i$ are the new coordinates. Because all the λ_i are strictly positive, then the set of vectors satisfying the latter equality is an ellipsoid centered at the origin with axes corresponding to the eigenvectors. The lengths of these axes are proportional to $1/\sqrt{\lambda_i}$. Hence, the larger the eigenvalue, the narrower is the ellipsoid in that direction. Figure A.3 illustrates this set when $n = 2$.

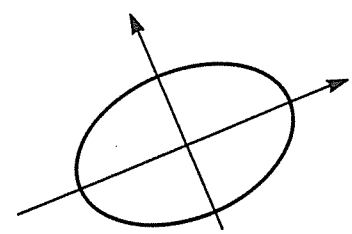


Figure A.3: Set of Vectors Satisfying $x'\Omega x = 1$

The vectors in the interior of such an ellipsoid are those for which $x'\Omega x \leq 1$. The volume of the ellipsoid is proportional to $\prod_{i=1}^n (1/\sqrt{\lambda_i})$, i.e., to $1/\sqrt{\det \Omega}$.

Now, consider two symmetric positive definite matrices such that $\Omega_1 \succeq \Omega_2$. This order relation is equivalent to $x'\Omega_1 x \geq x'\Omega_2 x, \forall x \in E$. Alternatively, because the above inequality is defined up to a positive multiplicative constant, then this inequality is also equivalent to $x'\Omega_1 x = 1 \Rightarrow 1 \geq x'\Omega_2 x$. Hence every point belonging to the ellipsoid associated with Ω_1 belongs to the interior of the ellipsoid associated with Ω_2. See Figure A.4. In particular, the volume of the first ellipsoid is smaller than or equal to the volume of the second ellipsoid. Thus $\det \Omega_1 \geq \det \Omega_2$. This provides a geometric proof of Corollary A.2.

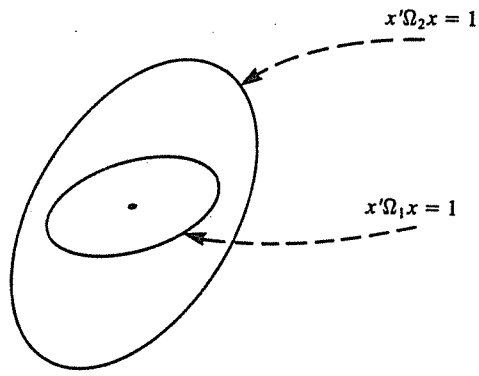


Figure A.4: Two Ellipsoids

A.5 Projections

Recall that two subspaces F and G of the vector space E are supplementary if

(i) $F \cap G = \{0\}$,

(ii) $E = F + G$.

Equivalently, every vector x of E can be decomposed uniquely as a sum of a vector y of F and a vector z of G. Moreover, as noted in Section A.2.3, when E is endowed with a scalar product, then F and $G = F^{\perp}$ are examples of supplementary spaces.

A.5.1 Definition and Characterization

Definition A.4: *The projection on F along G is the linear mapping denoted $\mathbf{P}_{F\|G}$ that associates to every vector x the element y of F in its decomposition on F and G. Then y is called the projection of x on F along G.*

Definition A.5: *In the special case where $G = F^{\perp}$, the projection on F along G is called the orthogonal projection on F. It is denoted $\mathbf{P}_F$.*

The mappings thus defined are linear mappings from E to E. See Figure A.5.

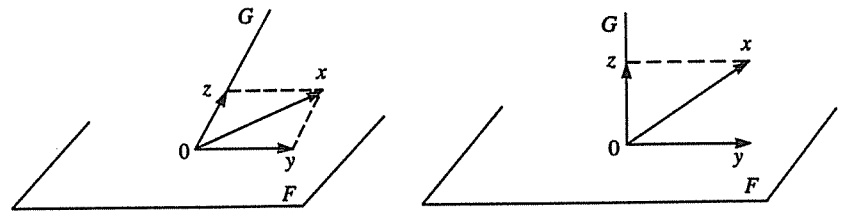


Figure A.5: Projections

The next property is obvious.

Property A.11: *We have $\mathbf{P}_{G\|F} = \mathbf{I} - \mathbf{P}_{F\|G}$. In particular, $\mathbf{P}_{F^{\perp}} = \mathbf{I} - \mathbf{P}_F$.*

Property A.12: *The following three statements are equivalent:*

(i) The linear mapping f is a projection on F along G.

(ii) The linear mapping f is idempotent, i.e., $f^2 = f$.

(iii) The linear mapping f can be diagonalized and its eigenvalues are either equal to one or zero.

If one of these conditions holds, then $F = \operatorname{Im} f$ and $G = \ker f$.

PROOF: $(i) \Rightarrow (iii)$: A vector y of F can be decomposed as $y = y + 0$, where $y \in F$ and $0 \in G$. Hence $f(y) = y$. Therefore, every vector of F is an eigenvector of f associated with the eigenvalue one. Similarly, every vector z of G can be written $z = 0 + z$, where $0 \in F$ and $z \in G$. Hence $f(z) = 0$. Thus every vector of G is an eigenvector of f associated with

the eigenvalue zero. Because $\dim F + \dim G = n$, it follows that f can be diagonalized.

(iii) $\Rightarrow$ *(ii)*: the linear mapping f can be written in matrix form as

$$\mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{Q}, \text{ with } \mathbf{\Delta} = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ddots \\ 0 & & & & 0 \end{pmatrix}.$$

Because $(\mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{Q})^2 = \mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{\Delta}^2\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{Q}$, it follows that the mapping f is idempotent.

(ii) $\Rightarrow$ *(i)*: Every vector x of E can be written as $y + z$, where $y = f(x) \in \text{Im } f$ and $z = x - f(x) \in \ker f$. Hence $\text{Im } f + \ker f = E$. It remains to show that $\ker f \cap \text{Im } f = \{0\}$.

Let $y_0 \in \ker f \cap \text{Im } f$. Because $y_0 \in \text{Im } f$, there exists a vector x_0 such that $y_0 = f(x_0)$. Moreover, because y_0 belongs to $\ker f$, we have

$$0 = f(y_0) = f(f(x_0)) = f(x_0) = y_0.$$

□

Property A.13: *The following three statements are equivalent:*

(i) The linear mapping f is an orthogonal projection on F.

(ii) The linear mapping f is idempotent and symmetric.

(iii) The linear mapping f can be diagonalized with respect to an orthonormal basis of eigenvectors and its eigenvalues are either equal to one or zero.

If one of these condition holds, then $F = \text{Im } f = \ker f^{\perp}$.

PROOF: *(i)* $\Rightarrow$ *(iii)*: The two eigenspaces F and $G = F^{\perp}$ are mutually orthogonal. Thus its is possible to choose an orthonormal basis that is composed of eigenvectors only.

(iii) $\Rightarrow$ *(ii)*: The linear mapping f can be written in matrix form as $\mathbf{Q}^{-1}\mathbf{\Delta}\mathbf{Q} = \mathbf{Q}'\mathbf{\Delta}\mathbf{Q}$ since we can choose $\mathbf{Q}$ to be orthogonal. Thus f is idempotent. The symmetry follows immediately.

(ii) $\Rightarrow$ *(i)*: Lastly, the orthogonality between Im f and $\ker f$ follows from Property A.6 and the symmetry of f. □

Corollary A.4: *If f is a projection, then* $\dim \text{Im}\ f = \text{rank}\ f = \text{tr}\ f$.

PROOF: The sum of the eigenvalues of f is equal to the number of eigenvalues equal to one, i.e., to the dimension of Im f. □

A.5.2 Matrix Representations of Projections

Property A.14: *Let $X_1, \ldots, X_K \in \mathbb{R}^n$ constitute a basis of the subspace F. The matrix representation of the orthogonal projection on F is $\mathbf{P}_F = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, where $\mathbf{X} = (X_1 \ldots X_K)$.*

PROOF: First, we note that $\mathbf{X}'\mathbf{X}$ is a symmetric matrix of size $K \times K$. Moreover, for every vector $b \in \mathbb{R}^K$ we have $b'\mathbf{X}'\mathbf{X}b = \|\mathbf{X}b\|^2$. Thus the quadratic form associated with the matrix $\mathbf{X}'\mathbf{X}$ is positive semidefinite. In addition, $b'\mathbf{X}'\mathbf{X}b = 0$ if and only if b is such that $\|\mathbf{X}b\|^2 = 0$, which is equivalent to $\mathbf{X}b = 0$. But, because the column vectors of $\mathbf{X}$ are linearly independent, this is equivalent to $b = 0$. Hence the matrix $\mathbf{X}'\mathbf{X}$ is positive definite.

Next, to verify that $\mathbf{P}_F$, as defined in Property A.14, is the orthogonal projection on F, it suffices to examine how $\mathbf{P}_F$ maps any element of F and $F^\perp$. We have:

(i) $Y \in F \Leftrightarrow \exists\ b \in \mathbb{R}^K : Y = \mathbf{X}b$. Hence $\mathbf{P}_F Y = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}b = \mathbf{X}b = Y$.

(ii) $Y \in F^\perp \Leftrightarrow \forall\ k = 1, \ldots, K, X_k'Y = 0 \Leftrightarrow \mathbf{X}'Y = 0$. Hence $\mathbf{P}_F Y = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = 0$.

Therefore F and $F^\perp$ are the eigenspaces of $\mathbf{P}_F$ associated with the eigenvalues one and zero, respectively. This completes the proof. □

A similar matrix representation can be derived for a projection on F along G, where G is no longer orthogonal to F. To understand such a matrix representation, the following lemma is useful.

Lemma A.2: *Every basis of $E = \mathbb{R}^n$ is an orthonomal basis with respect to an appropriate scalar product.*

PROOF: Consider a basis $A_1, \ldots, A_n$. We need to show that there exists a symmetric positive definite matrix Ω such that

$$\begin{aligned} A_i'\Omega A_j &= 1 \text{ if } i = j, \\ &= 0 \text{ otherwise.} \end{aligned}$$

Let $\mathbf{A} = (A_1 \dots A_n)$. The preceding condition can be written as

$$\mathbf{A}'\mathbf{\Omega}\mathbf{A} = \mathbf{I},$$

which is clearly satisfied by defining

$$\mathbf{\Omega} = (\mathbf{A}\mathbf{A}')^{-1} = \mathbf{A}'^{-1}\mathbf{A}^{-1}.$$

□

Note that

$$\begin{aligned} \tilde{\mathbf{A}}_i'\tilde{\mathbf{A}}_j &= 1 \text{ if } i = j, \\ &= 0 \text{ otherwise,} \end{aligned}$$

where $\tilde{\mathbf{A}}_i = \mathbf{\Omega}^{1/2}A_i = (\mathbf{A}\mathbf{A}')^{-1/2}A_i$. Thus we can always let $\tilde{\mathbf{A}}_i = \mathbf{\Omega}^{1/2}A_i$ so that the basis vectors are orthonormal and the scalar product corresponds to the identity matrix $\mathbf{I}$ with respect to this basis.

Property A.15: *Any projection (not necessarily orthogonal) on the subspace F generated by the vectors $X_1, \dots, X_K$ can be represented by a matrix of the form*

$$\mathbf{P}_{F\|G} = \mathbf{X}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega},$$

where $\mathbf{\Omega}$ is symmetric positive definite and G is the set of vectors Y satisfying $\mathbf{X}'\mathbf{\Omega}Y = 0$.

PROOF: We know that there exists a change of basis defined by $\mathbf{\Omega}^{1/2}$ such that the subspaces F and G become orthogonal. Using this change of basis, the matrix representing the orthogonal projection is $\tilde{\mathbf{P}}_{F\|G} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'$, where $\tilde{\mathbf{X}} = \mathbf{\Omega}^{1/2}\mathbf{X}$. Because $\tilde{\mathbf{P}}_{F\|G} = \mathbf{\Omega}^{1/2}\mathbf{P}_{F\|G}\mathbf{\Omega}^{-1/2}$, it follows that $\mathbf{P}_{F\|G} = \mathbf{X}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}$.

Lastly, the form of G directly follows from the orthogonality of F and G with respect to the scalar product defined by $\mathbf{\Omega}$. □

A.5.3 Some Properties of Orthogonal Projections

Property A.16: *Let $\mathbf{P}_{F_0}$ and $\mathbf{P}_{F_1}$ be two orthogonal projections on F_0 and F_1, respectively. Then*

$$\mathbf{P}_{F_0}\mathbf{P}_{F_1} = 0 \Leftrightarrow F_0 \text{ and } F_1 \text{ orthogonal.}$$

PROOF: To prove necessity, let $Y \in F_1$. Then we have $\mathbf{P}_{F_0}\mathbf{P}_{F_1}Y = \mathbf{P}_{F_0}Y = 0$. Thus Y belongs to $\ker \mathbf{P}_{F_0} = F_0^{\perp}$. Let Y vary in F_1, then we obtain $F_1 \subset F_0^{\perp}$.

The sufficiency part is obvious. □

Corollary A.5: *If F_0 and F_1 are two orthogonal subspaces, then*

$$\mathbf{P}_{F_0+F_1} = \mathbf{P}_{F_0} + \mathbf{P}_{F_1}.$$

PROOF: Every vector x can be decomposed uniquely as $x = x_0 + x_1 + x_2$, where $x_0 \in F_0$, $x_1 \in F_1$, and $x_2 \in (F_0 + F_1)^{\perp}$. Then it is easy to verify that $\mathbf{P}_{F_0+F_1}x = x_0 + x_1$, $\mathbf{P}_{F_0}x = x_0$, and $\mathbf{P}_{F_1}x = x_1$. This establishes the desired result. □

Corollary A.6: Frisch–Waugh Theorem. *We have*

$$\mathbf{P}_{F_0+F_1} = \mathbf{P}_{F_0} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{F_0})F_1}.$$

PROOF: The basic idea consists of using the preceding corollary by introducing two orthogonal subspaces. Let $F_1^* = (\mathbf{I} - \mathbf{P}_{F_0})F_1$ be the projection of F_1 on $F_0^{\perp}$. Then we have $F_0 + F_1^* = F_0 + F_1$. Moreover, the subspaces F_1^* and F_0 are orthogonal. It suffices now to apply the preceding corollary. □

The next property deals with the iterated projections.

Property A.17: *If $F_0 \subset F$, then $\mathbf{P}_{F_0}\mathbf{P}_F = \mathbf{P}_F\mathbf{P}_{F_0} = \mathbf{P}_{F_0}$.*

PROOF: Every vector x can be decomposed as $x = y + z + t$, where $y \in F_0$, z belongs to the subspace orthogonal to F_0 in F, and $t \in F^{\perp}$. Thus $\mathbf{P}_{F_0}x = y$ and $\mathbf{P}_F x = y + z$. Hence we clearly have

$$\mathbf{P}_{F_0}\mathbf{P}_F x = \mathbf{P}_F\mathbf{P}_{F_0}x = \mathbf{P}_{F_0}x = y.$$

□

Property A.17 says that projecting x first on F and then on F_0 is equivalent to projecting x directly on F_0. This is illustrated in Figure A.6.

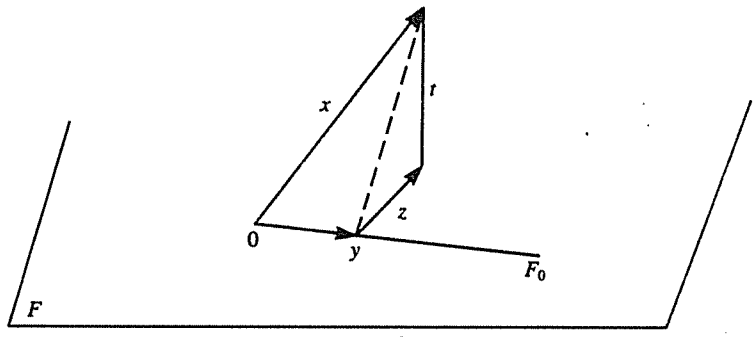


Figure A.6: Iterated Projections

A.5.4 An Optimal Property of Orthogonal Projections

Property A.18: *Let F be a subspace of a vector space E. Let x be a vector of E. Then there exists a unique element of F that is closest to x. This element is $\mathbf{P}_F x$, i.e., we have*

$$\|x - \mathbf{P}_F x\|^2 = \min_{z \in F} \|x - z\|^2.$$

PROOF: Let $z \in F$. We have

$$\begin{aligned} \|x - z\|^2 &= \|x - \mathbf{P}_F x + \mathbf{P}_F x - z\|^2 \\ &= \|x - \mathbf{P}_F x\|^2 + \|\mathbf{P}_F x - z\|^2 + 2 < x - \mathbf{P}_F x, \mathbf{P}_F x - z > \\ &= \|x - \mathbf{P}_F x\|^2 + \|\mathbf{P}_F x - z\|^2, \end{aligned}$$

because $x - \mathbf{P}_F x$ is equal to $\mathbf{P}_{F^\perp} x$ and, hence, is orthogonal to $\mathbf{P}_F x - z$.

It follows that $\forall\, z \in F : \|x - z\|^2 \geq \|x - \mathbf{P}_F x\|^2$. Hence $\|x - \mathbf{P}_F x\|^2 = \min_{z \in F} \|x - z\|^2$.

On the other hand, if z corresponds to the minimum, then we have necessarily $\|\mathbf{P}_F x - z\|^2 = 0$. Hence $z = \mathbf{P}_F x$, which shows that the minimum is attained uniquely at $P_F x$. □

A.6 Generalized Inverses

A.6.1 Definition and Construction

A linear mapping f from $E = \mathbb{R}^p$ to $F = \mathbb{R}^n$ is singular, in general. Its rank r, which is equal to the dimension of $\text{Im } f \subset F$, is an integer that is smaller than or equal to $\min\{p, q\}$. Its kernel $\ker f \subset E$ is of dimension $p - r$.

Thus the inverse of f does not exist in general. Nonetheless, this section extends the concept of the inverse of a linear mapping to the singular case.

Definition A.6: *A generalized inverse of f, denoted f^-, is a linear mapping from F to E satisfying*

$$\forall\, y \in \text{Im } f : f^-(y) \in f^{-1}(\{y\}) = \{x \in E : f(x) = y\}.$$

Recall that a linear mapping from F to E is defined by the images by this mapping of a basis of F. Now consider the decomposition of F into the direct sum of $\text{Im } f$ and $\text{Im } f^\perp$. Then suppose that a generalized inverse as defined above exists. Let $E_0 = f^-(\text{Im } f)$. If y_1 and y_2 are two distinct vectors of $\text{Im } f$, then their inverse images $f^{-1}(\{y_1\})$ and $f^{-1}(\{y_2\})$ do not intersect. Thus the mapping f^- is injective from $\text{Im } f$ to E_0. It follows that $E_0 = \dim \text{Im } f = r$. Moreover, it is easy to see that $E_0 \cap \ker f = \{0\}$. For, if x is an element of such an intersection, then $f(x) = 0$. But, because any linear mapping maps zero to zero, then we have necessarily $x = f^-(0) = 0$. Therefore E_0, which is equal to $f^-(\text{Im } f)$ and $\ker f$ are two supplementary spaces of E.

The preceding remarks allow us to explain how to construct generalized inverses of f.

(i) Choose a subspace E_0 of dimension r such that E_0 and $\ker f$ are supplementary spaces of E. Let $u_1, \ldots, u_r$ be a basis of E_0. Then $v_1 = f(u_1), \ldots, v_r = f(u_r)$ is a basis of $\text{Im } f$.

(ii) Then define the linear mapping f^- on $\text{Im } f$ by $f^-(v_j) = u_j$, $\forall\, j = 1, \ldots, r$. On the other hand, on $(\text{Im } f)^\perp$, f^- can be defined arbitrarily.

It is clear that there exists in general an infinity of generalized inverses. However, there exists a unique generalized inverse if $(\text{Im } f)^\perp = \{0\}$ and either $E_0 = \{0\}$ or $E_0 = \{E\}$, i.e., if $F = \text{Im } f$ and either

$\ker f = E$ or $\ker f = \{0\}$. The first case corresponds to the case where the linear mapping f is the zero mapping from E to $F = \{0\}$. The second case corresponds to the case where the linear mapping f is bijective from E to F.

Among the many generalized inverses, some may appear more intuitive. For instance, we can choose the space E_0 to be the subspace orthogonal to $\ker f$, and we can set f^- identically equal to zero on $(\text{Im } f)^\perp$. See Figure A.7. This particular generalized inverse is called the *pseudo inverse* or *Moore–Penrose inverse.* It is denoted f^+.

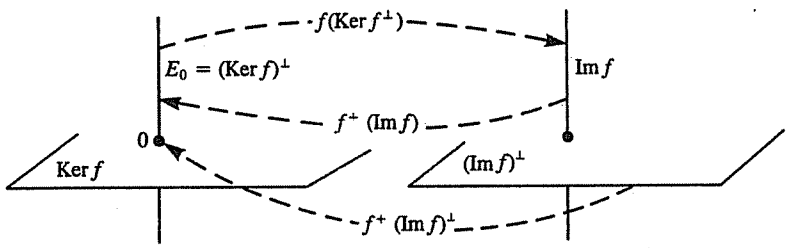


Figure A.7: Generalized Inverses

Property A.19: *Every linear mapping admits a generalized inverse and, more generally, an infinity of generalized inverses. With the exception of a degenerated case, the generalized inverse is unique if and only if f is nonsingular, in which case it coincides with f^{-1}.*

Property A.20: *The mapping f^- is a generalized inverse of the mapping f if and only if $ff^-f = f$.*

PROOF: We have

$$\begin{aligned} &\forall\, y \in \text{Im } f : f^-(y) \in f^{-1}(\{y\}) \\ &\quad \Leftrightarrow \quad \forall\, y \text{ of the form } f(x) : ff^-(y) = y \\ &\quad \Leftrightarrow \quad \forall\, x \in E : ff^-f(x) = f(x) \\ &\quad \Leftrightarrow \quad ff^-f = f. \end{aligned}$$

□

Note that we could have used Property A.20 to define a generalized inverse.

Property A.21: *Let f^- be a generalized inverse of f. Then every generalized inverse of f is of the form*

$$f^- + g\mathbf{P}_{(\mathrm{Im}\ f)^\perp} + \mathbf{P}_{\ker f}h,$$

where $\mathbf{P}_{(\mathrm{Im}\ f)^\perp}$ is the orthogonal projection on $(\mathrm{Im}f)^\perp$ in F, $\mathbf{P}_{\ker f}$ is the orthogonal projection on $\ker f$ in E, and g and h are some linear mappings from F to E.

PROOF: To prove sufficiency, it suffices to verify that

$$f(g\mathbf{P}_{(\mathrm{Im}\ f)^\perp} + \mathbf{P}_{\ker f}h)f = 0.$$

This follows immediately from $\mathbf{P}_{(\mathrm{Im}\ f)^\perp}f = 0$ and $f\mathbf{P}_{\ker f} = 0$.

To prove necessity, we note that the difference between two generalized inverses of f is a mapping k satisfying $fkf = 0$. Then let y be a vector of F. This vector can be decomposed as $y = y_0 + y_1$, where $y_0 = \mathbf{P}_{\mathrm{Im}\ f}y$ and $y_1 = \mathbf{P}_{(\mathrm{Im}\ f)^\perp}y$.

Now we have $k(y) = k(y_0) + k(y_1) = k(y_0) + k\mathbf{P}_{(\mathrm{Im}\ f)^\perp}y$. Moreover, $fk(y_0) = 0$ because $y_0 \in \mathrm{Im}\ f$. Hence $k(y_0) \in \ker f$. Thus we can write

$$\begin{aligned} k(y) &= \mathbf{P}_{\ker f}k(y_0) + k\mathbf{P}_{(\mathrm{Im}\ f)^\perp}(y) \\ &= \left(\mathbf{P}_{\ker f}k\mathbf{P}_{\mathrm{Im}\ f} + k\mathbf{P}_{(\mathrm{Im}\ f)^\perp}\right)y. \end{aligned}$$

This establishes the desired result. □

Note that the proof does not actually require that the projection on $\ker f$ be orthogonal.

A.6.2 Invariance Properties

In practice it is important to know which properties are independent of the chosen generalized inverse. We state such properties using the matrix representation of linear mappings.

Property A.22: *Let $\mathbf{A}$ be a matrix of size $n \times p$. Let $\mathbf{Q}$ and $\mathbf{P}$ be two nonsingular square matrices of size $n \times n$ and $p \times p$, respectively. Then the generalized inverses of the matrix $\mathbf{QAP}$ are of the form $\mathbf{P}^{-1}\mathbf{A}^-\mathbf{Q}^{-1}$, where $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$.*

PROOF: This is a consequence of the definition of generalized inverses of linear mappings. In particular, it is a consequence of the invariance of this definition with respect to a change of basis in E and F. □

Property A.23: *Let* $\mathbf{A}^-$ *be a generalized inverse of* $\mathbf{A}$. *Then* $\mathbf{A}\mathbf{A}^-$ *is a projection on* Im $\mathbf{A}$.

PROOF: The linear mapping $\mathbf{A}\mathbf{A}^-$ is a projection. To see this, it suffices to post-multiply the equality $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$ by $\mathbf{A}^-$ so as to conclude that $\mathbf{A}\mathbf{A}^-$ is idempotent.

The vectors of Im $\mathbf{A}$ are eigenvectors of $\mathbf{A}\mathbf{A}^-$ associated with the eigenvalue one. For, if $y \in$ Im $\mathbf{A}$ so that $y = \mathbf{A}x$ for some x, then we have $\mathbf{A}\mathbf{A}^-y = \mathbf{A}\mathbf{A}^-\mathbf{A}x = \mathbf{A}x = y$.

Lastly, we note that Im $\mathbf{A}\mathbf{A}^- \subset$ Im $\mathbf{A}$. But from above Im $\mathbf{A} \subset$ Im $\mathbf{A}\mathbf{A}^-$. Hence Im $\mathbf{A}\mathbf{A}^-$ = Im $\mathbf{A}$. Therefore dim ker $\mathbf{A}\mathbf{A}^- = n -$ dim Im $\mathbf{A}$. Hence ker $\mathbf{A}\mathbf{A}^-$ and Im $\mathbf{A}$ are supplementary subspaces of F. □

Corollary A.7: *We have* $\text{rank}(\mathbf{A}\mathbf{A}^-) = \text{tr}(\mathbf{A}\mathbf{A}^-) = \text{rank } \mathbf{A}$.

Property A.24: *The linear mapping* $\mathbf{A}^-\mathbf{A}$ *is a projection.*

PROOF: The proof is similar to the proof of Property A.23. □

Property A.25: *Let* $\mathbf{A}$ *be a symmetric matrix. If* $y \in$ Im $\mathbf{A}$, *then* $y'\mathbf{A}^-y$ *does not depend on the choice of the generalized inverse* $\mathbf{A}^-$ *of* $\mathbf{A}$.

PROOF: If $y \in$ Im $\mathbf{A}$, then there exists a vector x such that $y = \mathbf{A}x$. Thus we have $y'\mathbf{A}^-y = x'\mathbf{A}\mathbf{A}^-\mathbf{A}x = x'\mathbf{A}x$. □

The next property generalizes Property A.14 to the case where the column vectors of $\mathbf{A}$ are no longer linearly independent.

Property A.26: *The matrix* $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ *does not depend on the choice of the generalized inverse* $(\mathbf{A}'\mathbf{A})^-$ *of* $\mathbf{A}'\mathbf{A}$. *It is the matrix associated with the orthogonal projection on* Im $\mathbf{A}$.

PROOF: We shall show directly that the above matrix corresponds to the orthogonal projection on Im $\mathbf{A}$. The fact that this matrix is independent of the choice of generalized inverse will be an immediate consequence.

We have $\ker(\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}') \supset \ker \mathbf{A}' = (\text{Im } \mathbf{A})^\perp$ by Property A.6.

On the other hand, let $y = \mathbf{A}x \in$ Im $\mathbf{A}$. We have

$$\mathbf{A}'\left(\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{A}x - \mathbf{A}x\right) = 0.$$

Hence $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{A}x - \mathbf{A}x \in \ker \mathbf{A}' = (\text{Im } \mathbf{A})^\perp$. Because the latter vector also belongs to Im $\mathbf{A}$, it follows that it must be equal to the zero vector. Hence we have proved

$$\forall\, y \in \text{Im } \mathbf{A}, \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'y = y.$$

Lastly, the eigenspace E_1 associated with the eigenvalue one includes the subspace Im $\mathbf{A}$, while the eigenspace E_0 associated with the eigenvalue zero includes the subspace $(\text{Im } \mathbf{A})^{\perp}$. Because Im $\mathbf{A}$ and $(\text{Im } \mathbf{A})^{\perp}$ are supplementary subspaces of F, it follows that $E_1 = \text{Im } \mathbf{A}$ and $E_0 = (\text{Im } \mathbf{A})^{\perp}$. □

Note that classical properties of inverses are not necessarily satisfied by generalized inverses. For instance,

(i) if $\mathbf{A}$ is symmetric, then $\mathbf{A}^-$ is not necesssarily symmetric,

(ii) $(\mathbf{A}^-)^- \neq \mathbf{A}$, in general,

(iii) $(\mathbf{AB})^- \neq \mathbf{B}^-\mathbf{A}^-$, in general,

(iv) If $\mathbf{A}_n$ is a sequence of matrices converging to $\mathbf{A}$, and if $\mathbf{A}_n^-$ is a sequence of generalized inverses of $\mathbf{A}_n$, then $\mathbf{A}_n^-$ does not converge necessarily to a generalized inverse of $\mathbf{A}$. It suffices to consider the sequence $\mathbf{A}_n = 1/n$ so that $\mathbf{A} = 0$. Then $\mathbf{A}_n^- = \mathbf{A}_n^{-1} = n$. But $\mathbf{A}_n^-$ diverges.

A.6.3 Solving Linear Systems of Equations

Generalized inverses are especially used to solve for the unknown matrix $\mathbf{X}$ in linear systems of the form

$$\mathbf{AX} = \mathbf{B},$$

where $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{B}$ are matrices of size $n \times p$, $p \times m$, and $n \times m$, respectively.

Property A.27: *The preceding linear system has at least a solution if and only if* $\mathbf{AA}^-\mathbf{B} = \mathbf{B}$.

PROOF: A necessary and sufficient condition for the above system to have a solution is that the column vectors of $\mathbf{B}$ belong to the subspace Im $\mathbf{A}$. In other words, the column vectors of $\mathbf{B}$ are invariant with respect to the projection on Im $\mathbf{A}$. From Property A.23, the latter condition is equivalent to $\mathbf{AA}^-\mathbf{B} = \mathbf{B}$. □

Property A.28: *Let* $\mathbf{A}^-$ *be a generalized inverse of* $\mathbf{A}$. *The solutions of* $\mathbf{AX} = \mathbf{B}$ *are* $\mathbf{X} = \mathbf{A}^-\mathbf{B} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{V}$, *where* $\mathbf{V}$ *is an arbitrary matrix.*

PROOF: If there exists a solution to the system $\mathbf{AX} = \mathbf{B}$, then $\mathbf{B}$ can be written $\mathbf{AX}_0$. Thus we have

$$\begin{aligned}\mathbf{AX} &= \mathbf{A}^-\mathbf{AB} + \mathbf{A}(\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{V} \\ &= \mathbf{AA}^-\mathbf{AX}_0 + (\mathbf{A} - \mathbf{AA}^-\mathbf{A})\mathbf{V} \\ &= \mathbf{AX}_0 + 0 = \mathbf{B}.\end{aligned}$$

Hence every $\mathbf{X}$ of the form given in Property A.28 is a solution.

Conversely, if $\mathbf{X}_0$ is a solution, then $\mathbf{X}_0$ can be written as

$$\mathbf{X}_0 = \mathbf{A}^-\mathbf{B} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{X}_0.$$

This completes the proof. □

APPENDIX B

Review of Probability

B.1 Review of Integration Theory

B.1.1 Random Experiments

A *random experiment* can be formalized as a triplet $(\Omega, \mathcal{A}, P)$, where:

(i) Ω is the set of all possible outcomes of the experiment,

(ii) $\mathcal{A}$ is a σ-algebra, i.e., a set of subsets called *events* of Ω containing the sets $\emptyset$ and Ω, and which is closed under countable union and complement operations,

(iii) P is a set mapping from $\mathcal{A}$ to $\mathbb{R}^+$ satisfying $P(\Omega) = 1$ and the property: if $\{A_n\}$ is a sequence of *mutually exclusive* events (i.e., disjoint events), and if $\sum_n A_n$ denotes their union, then $P\left(\sum_n A_n\right) = \sum_n P(A_n)$. The latter property is referred as the σ-additivity property, and P is called a *probability measure* or a *probability distribution.*

The triplet $(\Omega, \mathcal{A}, \mathcal{P})$ is called a *probability space.*

More generally, any pair $(\Omega, \mathcal{A})$ satisfying (i) and (ii) can be endowed with a *measure.* Such a measure, denoted μ, is defined as a set mapping from $\mathcal{A}$ to $\mathbb{R}^+$ satisfying the σ-additivity condition only.

Thus a probability measure is simply a measure that assigns to the set Ω a *total measure* or *total mass* $\mu(\Omega)$ equal to one.

Some Properties of a Probability Measure

From the σ-additivity property and the condition $P(\Omega) = 1$, we readily obtain the following properties

$$P(^cA) = 1 - P(A), \ \forall \ A \in \mathcal{A}, \tag{B.1}$$

where ${}^cA$ denotes the complement of the event A

$$P\left(\bigcup_n A_n\right) \leq \sum_n P(A_n),\ \forall\ A_n \in \mathcal{A}, \tag{B.2}$$

$$P\left(\bigcap_n A_n\right) \geq 1 - \sum_n P({}^cA_n),\ \forall\ A_n \in \mathcal{A}, \tag{B.3}$$

$$P\left(\lim_n \uparrow A_n\right) = \lim_n \uparrow P(A_n), \tag{B.4}$$

for every increasing sequence of events $\{A_n\}$, and

$$P\left(\lim_n \downarrow A_n\right) = \lim_n \downarrow P(A_n), \tag{B.5}$$

for every decreasing sequence of events $\{A_n\}$.

B.1.2 Characterization of a Probability Distribution or a Measure

a) Discrete Distributions

When the set Ω of possible outcomes is finite or countable, and when $\mathcal{A}$ coincides with the set of all subsets of Ω, then a measure defined on $(\Omega, \mathcal{A})$ is said to be *discrete.*

We have the following obvious property.

Property B.1: *A discrete measure is fully characterized by all its values on sets that reduce to one element each.*

The *elementary* values or probabilities $p_\omega = P(\{\omega\})$ of a discrete probability distribution, where $\omega \in \Omega$, satisfy

$$p_\omega \geq 0,\ \forall\ \omega \in \Omega \text{ and } \sum_{\omega \in \Omega} p_\omega = 1.$$

Conversely, we have the next result.

Property B.2: *Let $\{p_\omega, \omega \in \Omega\}$ be a countable family of nonnegative numbers whose sum is equal to one. Then there exists a unique probability distribution on Ω whose elementary probabilities are $p_\omega, \omega \in \Omega$.*

An example of a discrete measure is the *counting measure* on $I\!N$ that associates to each subset of $I\!N$ the number of its elements.

b) Probability Distributions on $\mathbb{R}$

The set $\Omega = \mathbb{R}$ is usually endowed with the smallest σ-algebra of events containing the intervals of $\mathbb{R}$. This σ-algebra is called the *Borel* σ-algebra.

Definition B.1: *The cumulative distribution function of a probability measure P defined on $\mathbb{R}$ is the function F from $\mathbb{R}$ to $[0,1]$ defined by $F(x) = P((-\infty, x]),\ \forall\ x \in \mathbb{R}$.*

A cumulative distribution function is right-continuous, nondecreasing, and such that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$.

Conversely, we have the next result.

Property B.3: *Every real scalar function that is nondecreasing, right-continuous, and that satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$, can be viewed as the cumulative distribution function of a probability measure on $\mathbb{R}$. Moreover, this probability measure is unique.*

More generally, any measure on $\mathbb{R}$ that takes a finite value on every bounded interval is fully characterized by the set of values $\mu(]a,b])$, where $a < b$, and $a, b \in \mathbb{R}$. In particular, it can proved that there exists a measure λ corresponding to the intuitive notion of length, i.e., such that $\lambda(]a,b]) = b - a$. This measure is called the *Lebesgue measure* on $\mathbb{R}$.

c) Probability Distributions on $\mathbb{R}^n$

As in the previous subsection, a probability measure on $\mathbb{R}^n$ is characterized by its cumulative distribution function F. This function is a mapping from $\mathbb{R}^n$ to $[0,1]$ defined by

$$F(x_1, \ldots, x_n) = P\left(\prod_{i=1}^{n}(-\infty, x_i]\right).$$

As above, the concept of a cumulative distribution function can be generalized to a measure. In particular, it can be shown that there exists a unique measure λ_n that corresponds to the intuitive notion of volume. This measure satisfies

$$\lambda_n\left(\prod_{i=1}^{n}]x_i, y_i]\right) = \prod_{i=1}^{n}(y_i - x_i).$$

It is called the *Lebesgue measure* on $\mathbb{R}^n$.

B.1.3 Integration with Respect to a Measure

A measure can be viewed as a mapping defined on the set of indicator functions of events. This mapping is called the *integral with respect to the measure* μ. It is denoted $\int_\Omega \cdot \, d\mu$ or $\int_\Omega \cdot \, \mu(d\omega)$, and is defined by

$$\forall A \in \mathcal{A} : \mathbb{I}_A \overset{\int_\Omega \cdot \, d\mu}{\longrightarrow} \int_\Omega \mathbb{I}_A d\mu = \mu(A),$$

where $\mathbb{I}_A$ is the indicator function of the set A.

While still satisfying the semi-linearity property and the equality of limits under increasing sequences, this integral mapping can be extended uniquely to nonnegative functions of the form

$$f = \lim_n \uparrow \sum_{i=1}^{I_n} \alpha_{in} \mathbb{I}_{A_{in}}, \quad \alpha_{in} \geq 0, A_{in} \in \mathcal{A},$$

Namely, the integral of f is defined as

$$\begin{aligned} \int_\Omega f d\mu &= \int_\Omega \lim_n \uparrow \sum_{i=1}^{I_n} \alpha_{in} \mathbb{I}_{A_{in}} d\mu \\ &= \lim_n \uparrow \sum_{i=1}^{I_n} \int_\Omega \alpha_{in} \mathbb{I}_{A_{in}} d\mu \\ &= \lim_n \uparrow \sum_{i=1}^{I_n} \alpha_{in} \mu(A_{in}). \end{aligned}$$

Next, one can extend further this integral concept to functions f of arbitrary signs, provided their positive parts $f^+ = \max(f, 0)$ and negative parts $f^- = \max(-f, 0)$ both have finite integrals. Such functions are said to be *integrable* (with respect to the measure μ). In order to obtain an extension that is still linear, it is necessary to define the integral of such functions f as

$$\int_\Omega f d\mu = \int_\Omega f^+ d\mu - \int_\Omega f^- d\mu.$$

In practice integrals are easily computed by using the following two properties.

Property B.4: *Suppose that the measure μ is discrete. Then*

$$\int_\Omega f d\mu = \sum_{\omega \in \Omega} f(\omega) \mu(\{\omega\}).$$

Property B.5: *If f is a function from $\mathbb{R}^n$ to $\mathbb{R}$ that is integrable in the Riemann sense, then its integral with respect to the Lebesgue measure λ_n is equal to its Riemann integral, i.e.*

$$\int_{\mathbb{R}^n} f d\lambda_n = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

B.1.4 Density Functions

Property B.6: *Let f be a nonnegative function integrable with respect to a measure μ, i.e., such that $\int_\Omega f d\mu < \infty$. Then the mapping ν, which associates to any event A the number $\nu(A) = \int_\Omega \mathbb{1}_A f d\mu$, is a measure on $(\Omega, \mathcal{A})$. The function f is called a density of the measure ν with respect to μ. It is denoted $f = d\nu/d\mu$. Alternatively, we write $\nu = f\mu$.*

Note that the measure ν remains unchanged when f is modified on a set of measure zero with respect to μ. Thus the density f of ν is not unique.

Any discrete probability distribution on $\mathbb{N}$ has a density with respect to the counting measure on $\mathbb{N}$. This density is the mapping that associates the elementary probability p_n to every integer $n \in \mathbb{N}$.

When ν has a density f, then integrals with respect to ν can be easily obtained from integrals with respect to μ by using the following formula

$$\int_\Omega g d\nu = \int_\Omega g f d\mu. \tag{B.6}$$

Definition B.2: *A probability distribution P on $\mathbb{R}^n$ is absolutely continuous if it has a density function with respect to the Lebesgue measure on $\mathbb{R}^n$, i.e., if*

$$P = f\lambda_n.$$

The density function of an absolutely continuous probability distribution on $\mathbb{R}^n$ can be derived from the cumulative distribution function, as the next property states.

Property B.7: *A probability distribution on $\mathbb{R}^n$ is absolutely continuous if and only if its cumulative distribution function is continuous and*

the set of points, where the partial derivative $\partial^n F(x_1, \ldots, x_n)/\partial x_1 \ldots \partial x_n$ *does not exist, is of zero Lebesgue measure. Then a density* f *is given by*

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}.$$

In particular, a density of an absolutely continuous probability distribution on $\mathbb{R}$ is given by $f(x) = dF(x)/dx$, for any point x where the derivative exists.

B.1.5 Random Variables

Consider two pairs $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$, where $\mathcal{A}$ and $\mathcal{B}$ are σ-algebrae of Ω and $\mathcal{X}$, respectively. A *random variable* X is a mapping from Ω to $\mathcal{X}$ such that $X^{-1}(\mathcal{B}) \in \mathcal{A}$, $\forall\ B \in \mathcal{B}$.

A *real* random variable is a random variable with values in $\mathbb{R}$, where $\mathbb{R}$ is endowed with its Borel σ-algebra. Then such mappings are fully characterized by the condition

$$\forall b \in \mathbb{R}, X^{-1}((-\infty, b]) = \{X \leq b\} \in \mathcal{A}.$$

When a probability distribution P is defined on (Ω, A), then the random variable X induces a probability distribution on $(\mathcal{X}, \mathcal{B})$.

Property B.8: *Let* $(\Omega, \mathcal{A}, \mathcal{P})$ *be a probability space and* X *a random variable from* $(\Omega, \mathcal{A})$ *to* $(\mathcal{X}, \mathcal{B})$. *Then the set function* P^X *defined by* $P^X(B) = P(X^{-1}(B))$, $\forall\ B \in \mathcal{B}$, *is a probability distribution on* $(\mathcal{X}, \mathcal{B})$. *It is called the probability distribution induced by* X *of* P, *or the image probability distribution of* P *by* X, *or simply the probability distribution of* X.

When X is a random variable that can take only a finite or countable number of values, then the probability distribution of X is discrete. Then X is said to be a discrete random variable.

When X is a real random variable, then P^X is a probability distribution on $\mathbb{R}$ that is fully characterized by its cumulative distribution function F_X, defined by $F_X(x) = P^X((-\infty, x])$ for every $x \in \mathbb{R}$.

Integrals with respect to the induced probability distribution can be computed using the so-called *transfer formula* given in the next property.

Property B.9: Transfer Formula. *Let* f *be a real scalar function defined on* $\mathcal{X}$ *that is integrable with respect to the probability distribution*

P^X. *Then*

$$\int_\Omega f \circ X dP = \int_{\mathcal{X}} f dP^X.$$

The latter value is usually denoted $Ef(X)$ and is called the expectation of the variable $f(X)$.

Previously, we noted that the probability distribution of X is fully characterized by the values of the integrals over intervals with respect to P^X, i.e., by $F_X(x)$ for every $x \in \mathbb{R}$. Other functions have the same property.

Property B.10: *The probability distribution of X is fully characterized by the values of the integrals*

$$E\exp(itX) = E(\cos tX) + iE(\sin tX),$$

where i denotes the square root of -1 and $t \in \mathbb{R}$.

The function ϕ_x defined by $\phi_x(t) = E\exp(itX)$ is called *the characteristic function* of X.

B.1.6 Properties of the Expectation of a Real Random Variable

The next properties follow directly from the definition of the integral with respect to a probability distribution:

- Linearity

$$E(aX + bY + c) = aEX + bEY + c, \tag{B.7}$$

- Positiveness

$$\begin{aligned} &\text{if } X \geq 0, \text{ then } EX \geq 0, \\ &\text{if } X \geq 0 \text{ and } EX = 0, \text{ then } P(X = 0) = 1, \end{aligned} \tag{B.8}$$

- Jensen or convexity inequality

$$\text{if } h \text{ is convex, then } E(h(X)) \geq h(EX). \tag{B.9}$$

B.1.7 Some Descriptive Measures of the Probability Distribution of a Real Random Variable

The characteristics of the probability distribution of a real random variable X can be summarized in various ways:

- An indication of the *average value* taken by the random variable X is given by the expectation EX.
- The *dispersion* of the variable around the preceding value is frequently measured by either the *variance*, which is defined as

$$VX = E(X - EX)^2 = E(X^2) - (EX)^2, \tag{B.10}$$

or the *standard deviation* defined as

$$\sigma_x = \sqrt{VX} = \sqrt{E(X - EX)^2}. \tag{B.11}$$

- Other descriptive measures are the *quantiles*. If the cumulative distribution function F_X is continuous and strictly increasing, then the *quantile function* is the inverse function F_X^{-1}. The α-th quantile is the value $F_X^{-1}(\alpha)$ such that $P(X \leq F_X^{-1}(\alpha)) = \alpha$. In particular, the 50%-th quantile is called the *median*.

The preceding descriptive measures have the following properties

$$VX \geq 0, \ \forall \ X, \tag{B.12}$$

$$V(aX + b) = a^2 VX, \ \forall \ a, b \in I\!R, \tag{B.13}$$

$$VX = 0 \Leftrightarrow P(X = EX) = 1, \tag{B.14}$$

and, if G is a continuous strictly increasing function, and if $F_X^{-1}(1/2)$ is the median of X, then

$$\text{median of } G(X) = G\left(F_X^{-1}(1/2)\right). \tag{B.15}$$

B.1.8 Some Probability Distributions on $I\!R$

In Tables B.1 and B.2 we have gathered some well-known probability distributions on $I\!R$. In particular, we have indicated the domain (i.e., the set of possible values), the elementary probabilities in the case of a discrete probability distribution, the density function in the case of an absolutely

continuous probability distribution with respect to the Lebesgue measure, the cumulative distribution function (when this function is useful in practice), and the mean and the variance of the probability distribution.

Other probability distributions will be introduced when studying vectors of random variables and the normal distribution.

B.1.9 Random Vectors and Matrices

A random *vector* or *matrix* is a vector or a matrix whose elements are real random varriables defined on the same space $(\Omega, \mathcal{A})$. Thus the probability distribution of a random matrix $\mathbf{X}$ of size $n \times K$ with elements X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, K$, is a probability distribution on $\mathbb{R}^{nK}$.

The probability distribution of a subset of the elements X_{ij} is called a *marginal distribution.*

The concept of *expectation* can be generalized to random vectors or matrices by defining

$$E(\mathbf{X}) = (E(X_{ij})). \tag{B.16}$$

That is, the expectation of a matrix $\mathbf{X}$ is the matrix whose elements are the expectations of the elements of $\mathbf{X}$. Such an expectation has the following properties:

- If $\mathbf{A}$ is a nonrandom matrix of size $p \times n$, then

$$E(\mathbf{AX}) = \mathbf{A}E(\mathbf{X}). \tag{B.17}$$

- If $\mathbf{B}$ is a nonrandom matrix of size $K \times q$, then

$$E(\mathbf{XB}) = E(\mathbf{X})\mathbf{B}. \tag{B.18}$$

- The expectation and the transposition commute, i.e.

$$E(\mathbf{X}') = (E\mathbf{X})'. \tag{B.19}$$

In particular, the expectation of a symmetric random matrix is a symmetric matrix, and the expectation of a symmetric positive semidefinite matrix is a symmetric positive semidefinite matrix.

- The expectation and the trace operator commute, i.e.

$$E(\text{tr } \mathbf{X}) = \text{tr}(E\mathbf{X}). \tag{B.20}$$

The preceding properties can be combined to obtain the next result.

Table B.1: Discrete Distributions

Distribution	Domain	Probability	Mean EX	Variance VX
Dirac Measure (or point mass)	$\{x\}$	$p_x = 1$	x	0
Bernoulli $B(1,p)$, $p \in [0,1]$	$\{0,1\}$	$p_1 = p$ $p_0 = 1-p = q$	p	pq
Binomial $B(n,p)$, $p \in [0,1]$	$\{0,1,\ldots,n\}$	$p_x = \binom{n}{x} p^x q^{n-x}$ with $q = 1-p$	np	npq
Poisson $\mathcal{P}(\lambda)$, $\lambda \in \mathbb{R}^+$	$\mathbb{N}$	$p_x = \exp(-\lambda)\frac{\lambda^x}{x!}$	λ	λ
Geometric	$\mathbb{N}^*$	$p_x = p(1-p)^{x-1}$	$1/p$	q/p^2
Uniform $\mathcal{U}_{\{1,\ldots,n\}}$	$\{1,\ldots,n\}$	$p_x = 1/n$	$(n+1)/2$	$(n^2-1)/12$

Table B.2: Continuous Distributions

Distribution	Domain	Density Function	Cumulative Distribution Function	Mean EX	Variance VX
Uniform on $[a,b]$ $\mathcal{U}_{[a,b]}$	$[a,b]$	$\mathbb{1}_{[a,b]}(x)$	$\begin{cases} 0, \text{if } x \leq a \\ \frac{x-a}{b-a}, \text{if } x \in [a,b] \\ 1, \text{if } x \geq b \end{cases}$	$(a+b)/2$	$(b-a)^2/12$
Exponential $\gamma(1)$	$\mathbb{R}^+$	$\mathbb{1}_{\mathbb{R}^+}(x)\exp(-x)$	$\mathbb{1}_{\mathbb{R}^+}(x)(1-\exp(-x))$	1	1
$\gamma(1,\theta), \theta \in \mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{1}_{\mathbb{R}^+}(x)\frac{1}{\theta}\exp\frac{-x}{\theta}$	$\mathbb{1}_{\mathbb{R}^+}(x)\left(1-\exp\frac{-x}{\theta}\right)$	θ	θ^2
Cauchy	$\mathbb{R}$	$\frac{1}{\pi}\frac{1}{1+x^2}$	$\frac{1}{\pi}\arctan x + \frac{1}{2}$	do not exist	

Table B.2: Continued

Logistic	$\mathbb{R}$	$\frac{\exp(-x)}{(1+\exp(-x))^2}$	$\frac{1}{1+\exp(-x)}$	0	$\pi^2/3$
Standard Normal $N(0,1)$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$	$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp-\frac{t^2}{2}dt$	0	1
Gamma $\gamma(\nu)$, $\nu>0$	$\mathbb{R}^+$	$\mathbb{I}_{\mathbb{R}^+}(x)\frac{x^{\nu-1}\exp(-x)}{\Gamma(\nu)}$, where $\Gamma(\nu)=\int_0^\infty x^{\nu-1}\exp(-x)dx$	no closed form	ν	ν
Pareto, $\alpha>0, A>0$	$[A,+\infty)$	$\frac{\alpha A^\alpha}{x^{\alpha+1}}\mathbb{I}_{x\geq A}$	$1-\left(\frac{A}{x}\right)^\alpha$	$\frac{\alpha A}{\alpha-1}$, if $\alpha>1$	$\frac{\alpha A^2}{(\alpha-2)(\alpha-1)^2}$, if $\alpha>2$

Property B.11: *Let X be a random vector of dimension $n \times 1$. Let $\mathbf{A}$ be a nonrandom matrix of size $n \times n$. Then*

$$E(X'\mathbf{A}X) = \text{tr}(\mathbf{A}E(XX')).$$

PROOF: We have

$$\begin{aligned} E(X'\mathbf{A}X) &= \text{tr } E(X'\mathbf{A}X) \\ &= E \text{ tr}(X'\mathbf{A}X) \\ &= E \text{ tr}(\mathbf{A}XX') \\ &= \text{tr } E(\mathbf{A}XX') \\ &= \text{tr}(\mathbf{A}E(XX')), \end{aligned}$$

where the successive equalities follow respectively from the fact that $X'\mathbf{A}X$ is a scalar, equation (B.20), Property A.2, equation (B.20), and equation (B.17). □

Similarly, we can generalize the concept of variance to random vectors.

Definition B.3: *Let X and Y be two random vectors of size $n \times 1$ and $p \times 1$, respectively.*

(i) The variance covariance matrix of X is

$$VX = E(X - EX)(X - EX)'.$$

(ii) The covariance matrix between X and Y is

$$\text{Cov}(X, Y) = E(X - EX)(Y - EY)'.$$

The variance covariance matrix of X is a square matrix of size $n \times n$ whose diagonal elements are the variances of the components of X, and whose off-diagonal elements are the covariances between two distinct components of X. That is, we have

$$VX = \begin{pmatrix} VX_1 & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & VX_2 & & \\ \vdots & & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \dots & VX_n \end{pmatrix}. \tag{B.21}$$

The elements VX_i are dispersion measures. The off-diagonal element $\text{Cov}(X_i, X_j)$ measures the strength of the (linear) relationship between X_i and X_j. These relationships can be equally summarized by the so-called *correlation matrix*

$$\mathbf{R}(X) = \begin{pmatrix} 1 & \rho(X_1, X_2) & \dots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho(X_n, X_1) & \dots & \dots & 1 \end{pmatrix}, \tag{B.22}$$

where

$$\rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{VX_i}\sqrt{VX_j}}.$$

The variance covariance matrix and the correlation matrix have the following properties

$$V(X) \text{ is a positive semidefinite matrix.} \tag{B.23}$$

In particular, we have the *Schwarz inequality*

$$(\text{Cov}(X_i, X_j))^2 \leq VX_i VX_j, \tag{B.24}$$

or equivalently

$$-1 \leq \rho(X_i, X_j) \leq 1. \tag{B.25}$$

Moreover, it can be shown that

$$P(X \in EX + \text{Im } VX) = 1. \tag{B.26}$$

The latter property implies that

$$VX \text{ is nonsingular if and only if there does not exist any affine relationship between the components of } X. \tag{B.27}$$

The next properties follow immediately from the definition of the variance covariance matrix

$$VX = E(XX') - (EX)(EX)', \tag{B.28}$$

$$\text{Cov}(X, Y) = E(XY') - (EX)(EY)', \tag{B.29}$$

$$\text{Cov}(X, X) = VX, \tag{B.30}$$

$$\text{Cov}(Y, X) = \text{Cov}(X, Y)', \tag{B.31}$$

$$V(X+Y) = VX + VY + \text{Cov}(X, Y) + \text{Cov}(Y, X), \tag{B.32}$$

provided the vectors X and Y have the same size,

$$\text{Cov}(\mathbf{A}X, \mathbf{B}Y) = \mathbf{A}\text{Cov}(X, Y)\mathbf{B}', \tag{B.33}$$

provided the nonrandom matrices $\mathbf{A}$ and $\mathbf{B}$ have compatible size. In particular, we have $V(\mathbf{A}X + b) = \mathbf{A}V(X)\mathbf{A}'$. We also have

$$\text{tr } VX = E(\|X - EX\|^2) = E((X - EX)'(X - EX)). \tag{B.34}$$

The transfer formula given in Property B.9 still hold when X is a random vector. Moreover, the characteristic function can be generalized to random vectors.

Definition B.4: *The characteristic function of the random vector* $\mathbf{X} = (X_1, \ldots, X_n)'$ *is the mapping* ϕ_x *from* $\mathbb{R}^n$ *to* $\mathbb{C}$ *defined by*

$$\phi_x(u) = E(\exp iu'\mathbf{X}) = E\left(\exp i \sum_{j=1}^{n} u_j X_j\right).$$

Theorem B.1: The One-To-One Theorem. *The probability distribution of* X *is completely characterized by its characteristic function.*

A similar result also holds for the function $\psi_x(u) = E(\exp u'X)$, provided the latter function is defined on a neighborhood of $u = 0$. The function ψ_x is called the *Laplace transform* or *moment generating function* of X.

Lastly, we mention the Jacobian Theorem.

Property B.12: *Suppose that* X *has a density* $\mathbb{1}_{\mathcal{X}} f_x$ *with respect to the Lebesgue measure on* $\mathbb{R}^n$. *Let* $Y = g(X)$, *where* g *is a bijective and bidifferentiable function from* $\mathcal{X}$ *to* $\mathcal{Y} = g(\mathcal{X}) \subset \mathbb{R}^n$. *Then* Y *has a density with respect to* λ_n, *which is* $\mathbb{1}_{\mathcal{Y}} f_x \circ g^{-1} |J(g^{-1})|$, *where* $J(g^{-1})$ *is the determinant of the Jacobian of the transformation* g^{-1}.

B.2 Independence and Conditioning

B.2.1 Conditional Probability Distributions

Consider two random vectors X and Y of respective size p and q with values in $\mathcal{X}$ and $\mathcal{Y}$. We can define the *joint* probability distribution $P^{X,Y}$ of (X,Y) as well as the marginal probability distributions P^X and P^Y of X and Y. In this section we are interested in the relationship between the marginal probability distribution P^X (say) and the joint probability distribution $P^{X,Y}$.

Definition B.5: *A transition probability from $\mathcal{X}$ to $\mathcal{Y}$ is a measurable family of probability distributions on $\mathcal{Y}$ indexed by $x \in \mathcal{X}$.*

A transition probability is denoted $P_x(dy)$. From this definition it follows that, for every given x, the mapping that associates the value $P_x(A)$ to every event A of $\mathcal{Y}$ is a probability distribution. On the other hand, the mapping that associates the value $P_x(A)$ to every $x \in \mathcal{X}$ is a nonnegative scalar function defined on $\mathcal{X}$ that is integrable with respect to any given probability distribution on $\mathcal{X}$.

Property B.13: *There exists a unique transition probability from $\mathcal{X}$ to $\mathcal{Y}$ satisfying*

$$\begin{aligned} Eg(X,Y) &= \int_{\mathcal{X}\times\mathcal{Y}} g(x,y) P^{X,Y}(dxdy) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} g(x,y) P_x(dy) \right) P^X(dx), \end{aligned}$$

for every nonnegative or integrable scalar function g defined on $\mathcal{X} \times \mathcal{Y}$. Such a transition probability is denoted $P_x(dy) = P^{Y|X=x}(dy)$. It is called the conditional probability distribution of Y given $X = x$.

The preceding property also holds if we restrict the function $g(x,y)$ to belong to the class of indicator functions $g(x,y) = \mathbb{1}_A(x)\mathbb{1}_B(y)$.

Moreover, Property B.13 can be written by reversing the role of X and Y. In this case we obtain the conditional probability distribution of X given $Y = y$, which is characterized by

$$Eg(X,Y) = \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} g(x,y) P^{X|Y=y}(dx) \right) P^Y(dy),$$

for every nonnegative or integrable function g.

Conditional probability distributions are frequently easy to determine by involving the next properties.

Property B.14: *When the probability distribution $P^{X,Y}$ is discrete with elementary probabilities $p(x,y)$, then we have:*

(i) The marginal probability distribution of X is discrete with elementary probabilities

$$P^X(x) = \sum_y p(x,y) \equiv p(x).$$

(ii) The conditional probability distribution of Y given $X = x$ is discrete with elementary probabilities

$$P^{Y|X=x}(y) = \frac{p(x,y)}{\sum_y p(x,y)} = \frac{p(x,y)}{p(x)}.$$

The elementary probability $P^{Y|X=x}(y)$ is also denoted $P(Y = y \mid X = x)$.

Property B.15: *When the probability distribution of $P^{X,Y}$ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^p \times \mathbb{R}^q$ with a density $f(x,y)$, then we have:*

(i) The marginal probability distribution of X is absolutely continuous on $\mathcal{X}$ with a density given by

$$f(x) = \int_{\mathcal{Y}} f(x,y)dy.$$

(ii) The conditional probability distribution of Y given $X = x$ is absolutely continuous on $\mathcal{Y}$ with a density given by

$$f(y \mid x) = \frac{f(x,y)}{f(x)}.$$

Note that the conditional probability distribution is determined similarly in the above two properties. The methods consist in determining the conditional density function as the ratio of the joint density function and the marginal density function.

Property B.16: *Let X be a random vector whose probability distribution P^X is absolutely continuous with a density $f(x)$ with respect to some measure μ. Let h be a function from $\mathcal{X}$ to $\mathcal{Y} \subset \mathbb{R}^q$.*

(i) The probability distribution of $Y = h(X)$ is absolutely continuous with a density $f_h(y)$ with respect to some measure that depends on μ and h but not on f.

(ii) The conditional probability distribution of X given $Y = y = h(x)$ is absolutely continuous with respect to an appropriately chosen measure depending on μ and h but not on f. Its density is of the form

$$f(x \mid y) = \frac{f(x)}{f_h(y)} = \frac{f(x)}{f_h(h(x))}.$$

The decomposition of a joint probability distribution into the product of a marginal probability distribution and a conditional probability distribution can be generalized to an arbitrary number of random vectors. Specifically, consider the case of three random vectors X_1, X_2, and X_3. We can apply the preceding decomposition to either $X = X_1$ and $Y = (X_2, X_3)$ or $X = (X_1, X_2)$ and $Y = X_3$. We obtain

$$\begin{aligned} & Eg(X_1, X_2, X_3) \\ &= \int_{\mathcal{X}_1} \left(\int_{\mathcal{X}_2 \times \mathcal{X}_3} g(x_1, x_2, x_3) P^{2,3|1}(dx_2 dx_3) \right) P^{X_1}(dx_1) \\ &= \int_{\mathcal{X}_1 \times \mathcal{X}_2} \left(\int_{\mathcal{X}_3} g(x_1, x_2, x_3) P^{3|1,2}(dx_3) \right) P^{X_1, X_2}(dx_1 dx_2), \end{aligned}$$

where $P^{2,3|1} = P^{X_2, X_3 | X_1 = x_1}$ and $P^{3|1,2} = P^{X_3 | X_1 = x_1, X_2 = x_2}$.

Next, decomposing the joint probability distributions $P^{X_2, X_3 | X_1 = x_1}$ and P^{X_1, X_2}, we obtain

$$\begin{aligned} & Eg(X_1, X_2, X_3) \\ &= \int_{\mathcal{X}_1} \left(\int_{\mathcal{X}_2} \left(\int_{\mathcal{X}_3} g(x_1, x_2, x_3) P^{(3|2)|1}(dx_3) \right) P^{2|1}(dx_2) \right) P^{X_1}(dx_1) \\ &= \int_{\mathcal{X}_1} \left(\int_{\mathcal{X}_2} \left(\int_{\mathcal{X}_3} g(x_1, x_2, x_3) P^{3|1,2}(dx_3) \right) P^{2|1}(dx_2) \right) P^{X_1}(dx_1), \end{aligned}$$

where $P^{(3|2)|1} = P^{(X_3 | X_2 = x_2) | X_1 = x_1}$, $P^{2|1} = P^{X_2 | X_1 = x_1}$, and $P^{3|1,2} = P^{X_3 | X_1 = x_1, X_2 = x_2}$.

Now, from the uniqueness of the decomposition it follows that

$$P^{(X_3 | X_2 = x_2) | X_1 = x_1} = P^{X_3 | X_2 = x_2, X_1 = x_1}. \tag{B.35}$$

Hence the joint probability distribution of (X_1, X_2, X_3) can be obtained through successive conditionings. Formally, we have

$$P^{X_1,X_2,X_3} = P^{X_1} P^{X_2|X_1} P^{X_3|X_1,X_2}. \tag{B.36}$$

By induction, the latter formula can be generalized to an arbitrary number of random vectors as

$$P^{X_1,\ldots,X_n} = P^{X_1} P^{X_2|X_1} \ldots P^{X_n|X_1,\ldots,X_{n-1}}. \tag{B.37}$$

B.2.2 Conditional Expectations

The conditional probability distribution of Y given $X = x$ defines a probability distribution on $\mathcal{Y}$. Thus integrals with respect to this measure can be defined.

Definition B.6: *Let $g(X,Y)$ be a scalar function of two random vectors. Provided it exists, the conditional expectation of $g(X,Y)$ given $X = x$ is*

$$E(g(X,Y) \mid X = x) = \int_{\mathcal{Y}} g(x,y) P^{Y|X=x}(dy).$$

This conditional expectation can be viewed as defining a scalar function on $\mathcal{X}$. Thus it can be composed with the random variable X.

Definition B.7: *The conditional expectation of $g(X,Y)$ given the random variable X is the random variable*

$$E(g(X,Y) \mid X) = E(g(X,Y) \mid X = \cdot) \circ X.$$

From Property B.13 defining the conditional probability distribution, it follows that

$$E(g(X,Y)) = EE(g(X,Y) \mid X), \tag{B.38}$$

i.e., the expectation is equal to the expectation of the conditional expectation.

It is interesting to rewrite the latter equality using functions of the form $g(X,Y) = g_1(X)g_2(Y)$, where $g_1(X)$ and $g_2(Y)$ are square integrable. A direct computation of the involved integrals shows that

$$E(g_1(x)g_2(Y) \mid X) = g_1(X)E(g_2(Y) \mid X).$$

Taking expectation of each side gives

$$E(g_1(X)g_2(Y)) = E(g_1(X)E(g_2(Y) \mid X)),$$

i.e.

$$E(g_1(X)(g_2(Y) - E(g_2(Y) \mid X))) = 0. \tag{B.39}$$

Now, the set of square integrable random vectors is a vector space that can be endowed with the scalar product $< Z_1, Z_2 > = E(Z_1 Z_2)$. This space is denoted L^2. Thus equation (B.39), which holds for every square integrable function $g_1(X)$, can be interpreted as follows: The function $E(g_2(Y) \mid X)$, which is a function of X, is such that $g_2(Y) - E(g_2(Y) \mid X)$ is orthogonal to every function of X. Thus $E(g_2(Y) \mid X)$ is the orthogonal projection of $g_2(Y)$ on the set of square integrable functions of X. This is formally stated in the next property.

Property B.17: *The conditional expectation $E(g_2(Y) \mid X)$ is the orthogonal projection in L^2 of $g_2(Y)$ on the space $L^2(X)$ of square integrable functions of X.*

B.2.3 Properties of the Conditional Expectation

The properties of the conditional expectation follow from its interpretation as an orthogonal projection and as an expectation with respect to the conditional probability distribution. Before stating its properties, we note that the concept of conditional expectation can be applied to the vector Y itself.

Definition B.8: *The conditional expectation of the random vector Y of size q given X is the random vector of size q whose components are the conditional expectations of the components of Y, i.e.*

$$E(Y \mid X) = \begin{pmatrix} E(Y_1 \mid X) \\ \vdots \\ E(Y_q \mid X) \end{pmatrix}.$$

The main properties of $E(\cdot \mid X)$ are:

• Linearity: Let $\mathbf{A}$ and $\mathbf{B}$ be two nonrandom matrices. Let b be a nonrandom vector. Let Y_1 and Y_2 be two random vectors. Then,

provided the sizes of these matrices and vectors are compatible, we have

$$\begin{aligned} E(\mathbf{A}Y+b \mid X) &= \mathbf{A}E(Y \mid X)+b, \\ E(Y\mathbf{B}+b \mid X) &= E(Y \mid X)\mathbf{B}+b, \\ E(Y_1+Y_2 \mid X) &= E(Y_1 \mid X)+E(Y_2 \mid X). \end{aligned} \tag{B.40}$$

- Positiveness

$$\text{if } Y_i \geq 0,\ \forall\, i=1,\ldots,q, \text{then } E(Y_i \mid X) \geq 0,\ \forall\, i=1,\ldots,q. \tag{B.41}$$

- Monotony

$$\text{if } Y_i \geq Z_i,\ \forall\, i=1,\ldots,q, \text{then } E(Y_i \mid X) \geq E(Z_i \mid X),\ \forall\, i. \tag{B.42}$$

- The elements of $L^2(X)$, i.e., the square integrable functions of X, are invariant with respect to the operator $E(\cdot \mid X)$, i.e.

$$\forall\, i, Y_i \in L^2(X) \Leftrightarrow E(Y \mid X) = Y. \tag{B.43}$$

- The orthogonal projection of Y_i is the element of $L^2(X)$ that is closest to Y_i. It is given by

$$E(Y_i - E(Y_i \mid X))^2 = \min_{g_i(X)\in L^2(X)} E(Y_i - g_i(X))^2. \tag{B.44}$$

- Iterated Conditionings or Three-Perpendicular Theorem

$$\begin{aligned} &\text{if } L^2(Z) \subset L^2(X), \\ &\text{then } E(E(Y \mid X) \mid Z) = E(E(Y \mid Z) \mid X) = E(Y \mid Z). \end{aligned} \tag{B.45}$$

- Jensen or Convexity Inequality: If ψ is a convex function from $\mathbb{R}^q$ to $\mathbb{R}$, then

$$E(\psi(Y) \mid X) \geq \psi(E(Y \mid X)). \tag{B.46}$$

More generally, we can define the *conditional expectation of a random matrix* as the random matrix whose elements are the conditional expectations of the elements of the original matrix. This allows us to define the concept of conditional variance.

Definition B.9: *Let Y be a random vector of size q. The conditional variance covariance matrix of Y given X is the random square matrix of size q given by*

$$V(Y \mid X) = E((Y - E(Y \mid X))(Y - E(Y \mid X))' \mid X),$$

i.e., the matrix whose (i,j)-th element is

$$E((Y_i - E(Y_i \mid X))(Y_j - E(Y_j \mid X)) \mid X).$$

Similarly, we can define the conditional covariance matrix.

Definition B.10: *Let Y and Z be two random vectors. The conditional covariance matrix between Y and Z given X is the random matrix given by*

$$\mathrm{Cov}((Y,Z) \mid X) = E((Y - E(Y \mid X))(Z - E(Z \mid X))' \mid X).$$

We have the following properties

$$V(Y \mid X) = \mathrm{Cov}((Y,Y) \mid X), \tag{B.47}$$

$$\begin{aligned} \mathrm{Cov}((Y,Z) \mid X) &= E(YZ' \mid X) - E(Y \mid X)E(Z \mid X)', \\ V(Y \mid X) &= E(YY' \mid X) - E(Y \mid X)E(Y \mid X)'. \end{aligned} \tag{B.48}$$

If $\mathbf{A}$, $\mathbf{B}$, b, and c are nonrandom matrices and vectors of appropriate size, then

$$\begin{aligned} &\mathrm{Cov}((\mathbf{A}Y + b, \mathbf{B}Z + c) \mid X) = \mathbf{A}\ \mathrm{Cov}((Y,Z) \mid X)\ \mathbf{B}', \\ &V(\mathbf{A}Y + b \mid X) = \mathbf{A}\ V(Y \mid X)\ \mathbf{A}'. \end{aligned} \tag{B.49}$$

$V(Y \mid X)$ is a random symmetric positive semidefinite matrix. (B.50)

$$V\left(\begin{pmatrix} Y \\ Z \end{pmatrix} \mid X\right) = \begin{pmatrix} V(Y \mid X) & \mathrm{Cov}((Y,Z) \mid X) \\ \mathrm{Cov}((Z,Y) \mid X) & V(Z \mid X) \end{pmatrix}. \tag{B.51}$$

Another important property is the following one, which is analogous to equation (B.38).

Property B.18: Analysis of Variance. *We have*

$$VY = V(E(Y \mid X)) + EV(Y \mid X).$$

PROOF: We have

$$\begin{aligned} VY &= E(Y-E(Y))(Y-E(Y))' \\ &= E(Y-E(Y|X)+E(Y|X)-E(Y)) \\ &\quad (Y-E(Y|X)+E(Y|X)-E(Y))' \\ &= E(Y-E(Y|X))(Y-E(Y|X))' \\ &\quad + E(E(Y|X)-E(Y))(E(Y|X)-E(Y))', \end{aligned}$$

because of the orthogonality between $Y - E(Y \mid X)$ and any function of X. Then it suffices to note that

$$\begin{aligned} &E(Y-E(Y \mid X))(Y-E(Y \mid X))' \\ &\quad = EE((Y-E(Y|X))(Y-E(Y|X))'|X) \\ &\quad = EV(Y|X) \end{aligned}$$

and

$$\begin{aligned} &E(E(Y \mid X)-E(Y))(E(Y \mid X)-E(Y))' \\ &\quad = E(E(Y|X)-EE(Y|X))(E(Y|X)-EE(Y|X))' = VE(Y|X). \end{aligned}$$

□

B.2.4 Linear Regression

The conditional expectation $E(Y_i \mid X)$ is the L^2- orthogonal projection of Y_i on the subspace $L^2(X)$ spanned by the square integrable functions of X. Note that the latter functions can be linear or nonlinear in X. Sometimes we are interested in the orthogonal projection of Y_i on the linear (affine) functions of X. The subspace of linear functions of X is denoted

$$\mathcal{L}(X) = \{b_0 + X_1 b_1 + \ldots + X_p b_p, b_0, \ldots, b_p \in I\!R\}.$$

Definition B.11: *The linear (or affine) regression of the q-dimensional random vector Y on X is the random vector $EL(Y \mid X)$ of size q whose i-th component $EL(Y_i \mid X)$ is the orthogonal projection of Y_i on the space of linear functions of X.*

In general, the linear regression of Y on X differs from the conditional expectation of Y given X. As a matter of fact, because the subspace $\mathcal{L}(X)$ is a subspace of $L^2(X)$, the conditional expectation $E(Y \mid X)$

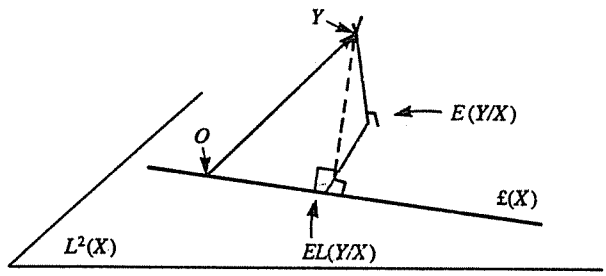


Figure B.1: Orthogonal Projections

always provides a better approximation of Y than $EL(Y \mid X)$. Moreover, from the Three-Perpendicular Theorem we have the following two properties

$$E\|Y - EL(Y \mid X)\|^2 \geq E\|Y - E(Y \mid X)\|^2 \tag{B.52}$$

and

$$EL(E(Y \mid X)) = EL(Y \mid X). \tag{B.53}$$

These two properties are illustrated in Figure B.1.

An advantage of the linear regression over the conditional expectation is that the former is simpler than the latter in view of the linearity of the elements of $\mathcal{L}(X)$.

Property B.19: *When the variance covariance matrix of X is nonsingular, then the linear regression of Y on X is given by*

$$EL(Y \mid X) = E(Y) + \mathrm{Cov}(Y, X)V(X)^{-1}(X - EX).$$

PROOF: The linear regression is of the form $EL(Y \mid X) = \mathbf{A}X + b$, where $\mathbf{A}$ and b are of size $q \times p$ and $q \times 1$, respectively. These quantities are determined by the orthogonality conditions

$$\begin{cases} E((Y - EL(Y \mid X)) \times 1) = 0, \\ E((Y - EL(Y \mid X)) \times X') = 0, \end{cases}$$

which are equivalent to

$$\begin{cases} E(Y) - \mathbf{A}E(X) - b = 0, \\ E(YX') - \mathbf{A}E(XX') - bE(X') = 0. \end{cases}$$

Hence we obtain

$$\begin{aligned} \mathbf{A} &= (E(YX') - EYEX')(E(XX') - EXEX')^{-1} \\ &= \mathrm{Cov}(Y,X)V(X)^{-1} \end{aligned}$$

and

$$b = EY - \mathbf{A}EX.$$

□

Definition B.12: *The residual (vector) of the linear regression is*

$$\epsilon L(Y \mid X) = Y - EL(Y \mid X).$$

Thus $\epsilon L(Y_i \mid X)$ can be interpreted as the orthogonal projection of Y_i on the subspace orthogonal to $\mathcal{L}(X)$.

Some properties of the mapping that associates its linear regression to the vector Y are similar to those of the conditional expectation. Namely, we have the linearity property

$$\begin{aligned} EL(\mathbf{A}Y + b \mid X) &= \mathbf{A}EL(Y \mid X) + b, \\ EL(Y_1 + Y_2 \mid X) &= EL(Y_1 \mid X) + EL(Y_2 \mid X). \end{aligned} \tag{B.54}$$

We also have

$$Y_i \in \mathcal{L}(X) \Leftrightarrow EL(Y_i \mid X) = Y_i, \tag{B.55}$$

and

$$EL(Y_i \mid X) = \hat{b}_0 + X_1\hat{b}_1 + \ldots + X_p\hat{b}_p, \tag{B.56}$$

where $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_p$ is a solution to

$$\min_{b_0, b_1, \ldots, b_p} E(Y_i - b_0 - X_1 b_1 - \ldots - X_p b_p)^2.$$

Theorem B.2: Iterated Regressions or Three-Perpendicular Theorem. *If $\mathcal{L}(Z) \subset \mathcal{L}(X)$, then*

$$EL(EL(Y \mid Z) \mid X) = EL(EL(Y \mid X) \mid Z) = EL(Y \mid Z).$$

On the other hand, the positiveness property (see equation (B.41)) and its consequences such as the monotony property and the convexity property (see equations (B.42) and (B.46)) are not satisfied by the linear regression mapping.

An important property is the following one.

Theorem B.3: Frisch–Waugh Theorem. *Let X, Y, and Z be three square integrable random vectors. We have*

$$EL(Y \mid X, Z) = EL(Y \mid X) + EL(Y \mid Z - EL(Z \mid X)) - EY.$$

PROOF: The space $\mathcal{L}(X, Z)$ can be decomposed as a direct sum of the two orthogonal subspaces $\mathcal{L}(X)$ and $\mathcal{L}^*(Z - EL(Z \mid X))$, where the latter subspace is the space spanned by the (nonaffine) linear combinations of the components of $Z - EL(Z \mid X)$. Because any element of this subspace is orthogonal to the unit vector, the projection of Y on this subspace is given by $EL(Y \mid Z - EL(Z \mid X)) - EY$.

The desired result follows by noting that the projection on (X, Z) is the sum of the projections on $\mathcal{L}(X)$ and $\mathcal{L}^*(Z - EL(Z \mid X))$ in view of the orthogonality of the latter two subspaces. □

Corollary B.1: *The coefficient matrix associated with Z in $EL(Y \mid X, Z)$ is equal to the coefficient matrix associated with $Z - EL(Z \mid X)$ in $EL(Y \mid Z - EL(Z \mid X))$.*

PROOF: We can write $EL(Y \mid X) = \mathbf{A}X + a$ and $EL(Y \mid Z - EL(Z \mid X)) = \mathbf{B}(Z - EL(Z \mid X)) + EY$. From Frisch–Waugh Theorem it follows that $EL(Y \mid X, Z) = \mathbf{A}X + a + \mathbf{B}(Z - EL(Z \mid X)) + EY$. Thus the coefficient of Z is equal to $\mathbf{B}$, as required. □

B.2.5 The Partial Variance Covariance Matrix

Definition B.13: *Let X, Y, and Z be three random vectors.*

(i) The partial variance covariance matrix or residual variance of Y given X is

$$VL(Y \mid X) = V(Y - EL(Y \mid X)) = V(\epsilon L(Y \mid X)).$$

(ii) The partial covariance matrix between Y and Z given X is

$$\mathrm{Cov}L(Y, Z \mid X) = \mathrm{Cov}(\epsilon L(Y \mid X), \epsilon L(Z \mid X)).$$

From the orthogonality conditions it follows at once that

$$\begin{aligned} VY &= VEL(Y \mid X) + V\epsilon L(Y \mid X) \\ &= VEL(Y \mid X) + VL(Y \mid X). \end{aligned} \tag{B.57}$$

This equation is called the *analysis-of-variance equation* or the *analysis-of-variance decomposition.* But

$$VEL(Y \mid X) = V(\mathrm{Cov}(Y, X)(VX)^{-1}X) = \mathrm{Cov}(Y, X)(VX)^{-1}\mathrm{Cov}(X, Y).$$

Hence the partial variance covariance matrix is given by

$$VL(Y \mid X) = VY - \mathrm{Cov}(Y, X)(VX)^{-1}\mathrm{Cov}(X, Y).$$

In particular, we obtain the following inequality.

Property B.20: Multivariate Cauchy–Schwarz Inequality.

$$VY \succeq \mathrm{Cov}(Y, X)(VX)^{-1}\mathrm{Cov}(X, Y).$$

The partial covariance matrix between Y and Z given X measures the strength of the (linear) relationship between Y and Z after having removed their linear dependencies on X.

A special case of interest arises when this partial covariance matrix is zero.

Definition B.14: *The random vectors Y and Z are linearly uncorrelated given X if* $\mathrm{Cov}(\epsilon L(Y \mid X), \epsilon L(Z \mid X)) = 0$. *Such a property is denoted* $Y \perp Z \mid X$.

Property B.21: *The following properties are all equivalent*

(i) $Y \perp Z \mid X$.

(ii) $Z \perp Y \mid X$.

(iii) $Y \perp (Z - EL(Z \mid X)) \mid X$.

(iv) $EL(Y \mid Z, X) = EL(Y \mid X)$.

(v) $EL(Y \mid Z - EL(Z \mid X)) = EY$.

(vi) $\mathrm{Cov}(Y, Z) = \mathrm{Cov}(Y, X)(VX)^{-1}\mathrm{Cov}(X, Z)$.

(vii) $VL(Y \mid Z, X) = VL(Y \mid X)$.

(viii) $Y_i \perp Z_j \mid X$ *for every pair of components* Y_i *and* Z_j *of* Y *and* Z.

PROOF: These equivalences are direct consequences of the definition of linear uncorrelation viewed as an orthogonality condition. □

Property B.22: *Let* Z_j, $j = 1, \ldots, J$, *be the components of the random vector* Z. *Then* Y *and* Z *are linearly uncorrelated given* X *if and only if*

$$\begin{aligned} Y &\perp Z_1 \mid X, \\ Y &\perp Z_2 \mid X, Z_1, \\ &\vdots \\ Y &\perp Z_J \mid X, Z_1, \ldots, Z_{J-1}. \end{aligned}$$

PROOF: To prove necessity, we note that the condition $Y \perp Z \mid X$ is equivalent to $EL(Y \mid Z, X) = EL(Y \mid X)$. Applying the operator $EL(\cdot \mid Z_1, \ldots, Z_{j-1}, X)$ to each side of this inequality, and invoking Theorem B.2 on iterated regressions, we obtain

$$EL(Y \mid Z_1, \ldots, Z_{j-1}, X) = EL(Y \mid X), \ \forall\, j.$$

This implies

$$EL(Y \mid Z_1, \ldots, Z_j, X) = EL(Y \mid Z_1, \ldots, Z_{j-1}, X), \ \forall\, j,$$

i.e.

$$Y \perp Z_j \mid Z_1, \ldots, Z_{j-1}, X, \ \forall\, j.$$

Conversely, if $EL(Y \mid Z_1, \ldots, Z_j, X) = EL(Y \mid Z_1, \ldots, Z_{j-1}, X)$, $\forall\, j$, then we have $EL(Y \mid Z_1, \ldots, Z_J, X) = EL(Y \mid X)$, i.e. $Y \perp Z \mid X$. □

B.2.6 Independence

A property that is stronger than the property of zero linear correlation is that of independence. In this subsection we review the main results concerning the concept of independence. In the next subsection we discuss the concept of conditional independence.

Definition B.15: *Two random vectors X and Y are (mutually) independent if*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B),$$

for every events A and B.

The preceding equality can be written under integral forms as

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Y}} \mathbb{I}_{A\times B}(x,y)P^{X,Y}(dx,dy) \\
&= \int_{\mathcal{X}} \mathbb{I}_A(x)P^X(dx) \int_{\mathcal{Y}} \mathbb{I}_B(y)P^Y(dy) \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \mathbb{I}_{A\times B}(x,y)P^Y(dy) \right) P^X(dx).
\end{aligned}$$

Thus, in view of Property B.13 the conditional probability distribution of Y given $X = x$ is identical to the marginal probability distribution of Y if Y and X are independent. By reversing the role of X and Y, we obtain a similar results for the conditional probability distribution of X given $Y = y$.

Property B.23: *We have*

(i) X and Y are independent if and only if

$$P^{Y|X=x} = P^Y, \ \forall\, x.$$

(ii) X and Y are independent if and only if

$$P^{X|Y=y} = P^X, \ \forall\, y.$$

These conditions can also be written in terms of densities. Specifically, from Property B.15 we obtain the next characterization.

Property B.24: *Suppose that X and Y are absolutely continuous with some densities $f_X(x)$ and $f_Y(y)$, respectively. If these random vectors are independent, then the joint distribution of (X,Y) is absolutely continuous with a density given by*

$$f(x,y) = f_X(x)f_Y(y).$$

That is, the joint density of the pair (X, Y) is just the product of the marginal densities.

The equality between the conditional density and the marginal density can be written alternatively in terms of expectations.

Property B.25: *The random vectors X and Y are independent if and only if $E(g(Y) \mid X) = Eg(Y)$ for every square integrable function g.*

Then, interpreting the conditional expectation as an orthogonal projection, we obtain immediately another characterization of independence from the preceding property.

Property B.26: *The random variables X and Y are independent if and only if $E(g(Y)h(X)) = Eg(Y)Eh(X)$ for every square integrable functions g and h.*

It is important to note that Properties B.25 and B.26 hold provided *all* square integrable functions g and h are considered. By considering only some specific functions, we readily obtain some necessary conditions for independence. In general, however, these conditions are not sufficient.

Property B.27: *Some necessary conditions for independence between X and Y are*

(i) $E(Y \mid X) = EY$.

(ii) $EL(Y \mid X) = EY$.

(iii) $\mathrm{Cov}(X, Y) = 0$.

(iv) If g and h are two functions, then $g(X)$ and $h(Y)$ are independent.

Note that additional characterizations of independence can be obtained by requiring that the equality of Property B.26 holds for some appropriately chosen classes of functions g and h. For instance, let $h(X) = \exp(iu'X)$ and $g(Y) = \exp(iv'Y)$, where u and v are vectors varying in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Such a choice leads naturally to characteristic functions.

Property B.28: *The random vectors X and Y are independent if and only if*

$$\phi_{X,Y}(u, v) = \phi_X(u)\phi_Y(v), \ \forall\, u \in \mathbb{R}^p, v \in \mathbb{R}^q,$$

i.e., if and only if

$$E \exp i(u'X + v'Y) = E(\exp iu'X)E(\exp iv'Y), \ \forall\, u, v.$$

The preceding definitions and properties can be easily generalized to an arbitrary number of random vectors. We have the following equivalences.

Property B.29: *The random vectors $X_1, \ldots, X_n$ are (mutually) independent if and only if either one of the following statements holds:*

(i)

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i), \ \forall \ A_i.$$

(ii)

$$P^{X_2|X_1} = P^{X_2}, P^{X_3|X_2,X_1} = P^{X_3}, \ldots, P^{X_n|X_{n-1},\ldots,X_1} = P^{X_n}.$$

(iii)

$$E\left(\prod_{i=1}^{n} g_i(X_i)\right) = \prod_{i=1}^{n} E(g_i(X_i)), \ \forall \ g_i.$$

(iv)

$$E(g_2(X_2) \mid X_1) = Eg_2(X_2), E(g_3(X_3) \mid X_2, X_1) = Eg_3(X_3),$$
$$\ldots, E(g_n(X_n) \mid X_{n-1}, \ldots, X_1) = Eg_n(X_n), \ \forall \ g_i.$$

(v)

$$\phi_{X_1,\ldots,X_n}(u_1, \ldots, u_n) = \prod_{i=1}^{n} \phi_{X_i}(u_i), \ \forall \ u_i.$$

B.2.7 Conditional Independence

Independence is a property of probability distributions. In particular, it applies to the conditional probability distribution of X and Y given Z.

Definition B.16: *The random vectors Y and Z are conditionally independent given X if*

$$P(Y \in A, Z \in B \mid X = x) = P(Y \in A \mid X = x)P(Z \in B \mid X = x),$$

for every events A and B and every value x. This conditional independence is denoted $Y \perp\!\!\!\perp Z \mid X$.

The results of the preceding subsection apply directly.

Property B.30: *The following statements are equivalent:*

(i) $Y \perp\!\!\!\perp Z \mid X$.

(ii) $P^{Y|Z,X} = P^{Y|X}$.

(iii) $E(g(Y)h(Z) \mid X) = E(g(Y) \mid X)E(h(Z) \mid X), \forall\, g, h.$

(iv) $E(g(Y) \mid Z, X) = E(g(Y) \mid X), \forall\, g.$

Though there is in general no relationship between partial uncorrelation and conditional independence, the later has some features that are analogous to those of a zero partial correlation. Hereafter we give some recursive characterizations of the conditional independence property. Let p, q, and J be the dimensions of the vectors X, Y, and Z, respectively.

Property B.31: *We have*

$$\begin{aligned}&(Y_1, \ldots, Y_q)' \perp\!\!\!\perp (Z_1, \ldots, Z_J)' \mid (X_1, \ldots, X_p)' \\ &\Leftrightarrow (Y_1^*, \ldots, Y_q^*) \perp\!\!\!\perp (Z_1^*, \ldots, Z_J^*)' \mid (X_1^*, \ldots, X_p^*)',\end{aligned}$$

where

$$\begin{aligned}X_i^* &= X_i - E(X_i \mid X_1, \ldots, X_{i-1}), \textit{with } X_1^* = X_1, \\ Y_i^* &= Y_i - E(Y_i \mid Y_1, \ldots, Y_{i-1}), \textit{with } Y_1^* = Y_1, \\ Z_i^* &= Z_i - E(Z_i \mid Z_1, \ldots, Z_{i-1}), \textit{with } Z_1^* = Z_1.\end{aligned}$$

PROOF: The random vectors X^*, Y^*, and Z^* are related to X, Y, and Z through some bijective mappings

$$X^* = h_1(X), \quad Y^* = h_2(Y), \quad Z^* = h_3(Z).$$

Therefore we have

$$\begin{aligned}&E(g_2(Y^*)g_3(Z^*) \mid X^* = x^*) \\ &= E(g_2(h_2(Y))g_3(h_3(Z)) \mid h_1(X) = x^*) \\ &= E(g_2(h_2(Y))g_3(h_3(Z)) \mid X = h_1^{-1}(x^*)) \\ &= E(g_2(h_2(Y)) \mid X = h_1^{-1}(x^*))E(g_3(h_3(Z)) \mid X = h_1^{-1}(x^*)) \\ &= E(g_2(Y^*) \mid X^* = x^*)E(g_3(Z^*) \mid X^* = x^*).\end{aligned}$$

Thus $X \perp\!\!\!\perp Y \mid Z \Rightarrow X^* \perp\!\!\!\perp Y^* \mid Z^*$.

The converse is established similarly by considering the inverse mappings h_1^{-1}, h_2^{-1}, and h_3^{-1}. □

Property B.32: *We have*

$$Y \perp\!\!\!\perp (Z_1, \ldots, Z_J)' \mid X \Leftrightarrow \begin{cases} Y \perp\!\!\!\perp Z_1 \mid X, \\ Y \perp\!\!\!\perp Z_2 \mid X, Z_1, \\ \vdots \\ Y \perp\!\!\!\perp Z_J \mid X, Z_1, \ldots, Z_{J-1}. \end{cases}$$

PROOF: The proof is similar to the proof of Property B.22 provided we replace $EL(Y \mid X, Z_1, \ldots, Z_{j-1})$ by $E(g_2(Y) \mid X, Z_1, \ldots, Z_{j-1})$ and we consider arbitrary functions g_2. □

B.3 Normal Distributions

B.3.1 Univariate Normal Distributions

Definition B.17: *The (univariate) standard normal distribution, which is denoted $N(0,1)$, is the distribution on $\mathbb{R}$, absolutely continuous with respect to the Lebesgue measure, with density*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Figure B.2 graphs the standard normal density. This density has a "bell shape" with two inflection points at ± 1.

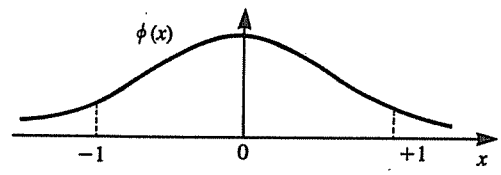


Figure B.2: Density of the Standard Normal Distribution

The cumulative probability distribution of the $N(0,1)$ distribution does not have a closed form. It is denoted $\Phi(\cdot)$ and is defined by

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt.$$

By differentiating and integrating by parts, we obtain the next property using the expressions for $\phi(\cdot)$ and $\Phi(\cdot)$.

Property B.33: *We have*

(i) $\phi(-x) = \phi(x)$.

(ii) $\Phi(-x) = 1 - \Phi(x)$. *Hence* $\Phi(0) = 1/2$.

(iii) $d\phi(x)/dx = -x\phi(x)$.

(iv) $\int_{-\infty}^{x} \Phi(t)dt = x\Phi(x) + \phi(x)$.

(v) $\phi(x)/\Phi(x) \# -x$ *when* $x \to -\infty$.

(vi) $\phi(x)/(1 - \Phi(x)) \# x$ *when* $x \to +\infty$.

The standard normal distribution has moments of any order. If $Z \sim N(0,1)$, then the moments are given by

$$\begin{aligned} E(Z^{2p+1}) &= 0, \ \forall\, p \in \mathbb{N}, \\ E(Z^{2p}) &= \frac{(2p)!}{2^p p!}, \ \forall\, p \in \mathbb{N}. \end{aligned} \tag{B.58}$$

In particular, if $Z \sim N(0,1)$, then

$$EZ = 0 \quad \text{and} \quad VZ = 1. \tag{B.59}$$

The normal distribution can be generalized to random variables with arbitrary means and variances.

Definition B.18: *A random variable X follows a normal distribution $N(m, \sigma^2)$ with mean m and variance σ^2 if $X = \sigma Z + m$, where Z follows a $N(0,1)$ distribution.*

When σ is nonzero, the random variable X follows an absolutely continuous distribution with density $(1/\sigma)\phi((x-m)/\sigma)$. Its cumulative probability distribution is $\Phi((x-m)/\sigma)$. Its mean and variance are m and σ^2, respectively. Its fourth centered moment, which is sometimes useful, is

$$E(X-m)^4 = 3\sigma^4 = 3(VX)^2. \tag{B.60}$$

The Laplace transform and the characteristic function of a normal distribution are quite simple.

Property B.34: *Let X be a random variable distributed as $N(m, \sigma^2)$. Then*

$$E(\exp tX) = \exp\left(tm + \frac{t^2\sigma^2}{2}\right)$$

and

$$\phi_X(t) = E(\exp itX) = \exp\left(itm - \frac{t^2\sigma^2}{2}\right).$$

The characteristic function can be used to show that the family of normal distributions is closed under linear combinations.

Property B.35: *Let X_i, $i = 1, \ldots, n$, be n independent random variables distributed as $N(m_i, \sigma_i^2)$, $i = 1, \ldots, n$, respectively, where a_i, $i = 1, \ldots, n$, and b are scalars. Then the random variable*

$$X = \sum_{i=1}^{n} a_i X_i + b$$

is distributed as $N(m, \sigma^2)$ with

$$m = \sum_{i=1}^{n} a_i m_i + b \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

Corollary B.2: *Let $X_1, \ldots, X_n$ be n independent random variables with identical distribution $N(m, \sigma^2)$. Then the sample mean of these variables $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is distributed as $N\left(m, \sigma^2/n\right)$.*

B.3.2 Multivariate Normal Distributions

a) Gaussian Random Vectors

Definition B.19: *A random vector X with values in $\mathbb{R}^p$ is Gaussian if every linear combination of its components $\alpha' X = \sum_{i=1}^{p} \alpha_i X_i$ follows a (possibly degenerated) normal distribution on $\mathbb{R}$.*

Note that we include the special case where the distribution $N(m, \sigma^2)$ of $\alpha' X$ is such that $\sigma^2 = 0$. In this special case the distribution $N(m, 0)$ coincides with the mass distribution at m.

Definition B.19 implies some important properties of Gaussian random vectors.

Property B.36: *Let X be a Gaussian random vector of size p. Let $\mathbf{A}$ and b be nonrandom matrices of size $q \times p$ and $q \times 1$, respectively. Then $Y = \mathbf{A}X + b$ is a Gaussian random vector. In particular, every subvector of a Gaussian random vector is Gaussian, and any of its components follows a normal distribution on $\mathbb{R}$.*

Property B.37: *Every Gaussian random vector has a finite mean $EX = m \in \mathbb{R}^p$ and a finite variance covariance matrix $VX = \mathbf{\Sigma}$.*

It remains to prove that there exist Gaussian random vectors for every possible choice of m and $\mathbf{\Sigma}$.

Property B.38: *Let m be a vector of $\mathbb{R}^p$ and $\mathbf{\Sigma}$ be a symmetric positive semidefinite matrix of size p. Then there exists a Gaussian random vector X with mean m and variance covariance matrix $\mathbf{\Sigma}$.*

PROOF: First, we note that there exists a Gaussian random vector with mean $m = 0$ and variance covariance matrix $\mathbf{\Sigma} = \mathbf{I}$. It suffices to take $Z = (Z_1, \ldots, Z_n)$, where the components Z_i are independent random variables distributed as $N(0, 1)$. Then the random vector Z has mean zero and variance covariance matrix equal to $\mathbf{I}$. Moreover, it is clearly Gaussian in the sense of Definition B.19.

Next, we define the vector $X = \mathbf{\Sigma}^{1/2} Z + m$. Then X has mean m and variance covariance matrix equal to $\mathbf{\Sigma}$. It is also Gaussian since Z is Gaussian. □

b) Multivariate Normal Distributions

Definition B.20: *A multivariate normal distribution on $\mathbb{R}^p$ is the probability distribution of a Gaussian vector of size p.*

Property B.39: *The characteristic function of a multivariate normal distribution with mean m and variance covariance matrix $\mathbf{\Sigma}$ is*

$$\phi_X(u) = E \exp(iu'X) = \exp\left(iu'm - \frac{u'\mathbf{\Sigma}u}{2}\right).$$

PROOF: Because X is a Gaussian random vector, then $u'X$ follows a normal distribution on $\mathbb{R}$. The desired result follows immediately from Property B.34. □

In particular, the multivariate normal distribution is fully characterized by ϕ_X, and therefore by its mean m and its variance covariance matrix $\mathbf{\Sigma}$. This is formally stated in the next property.

Property B.40: *A multivariate normal distribution is fully characterized by the knowledge of* m *and* Σ. *This distribution is denoted* $N(m, \Sigma)$.

Property B.40 can be used to determine easily the probability distribution of an arbitrary linear transformation of Gaussian vectors. For instance, we have

$$\text{if } X \sim N(m, \Sigma), \text{ then } \mathbf{A}X + b \sim N(\mathbf{A}m + b, \mathbf{A}\Sigma\mathbf{A}'), \tag{B.61}$$

$$\text{if } X \sim N(m, \Sigma) \text{ with } \Sigma \text{ nonsingular,}$$
$$\text{then } \Sigma^{-1/2}(X - m) \sim N(0, \mathbf{I}), \tag{B.62}$$

$$\text{if } X \sim N(0, \mathbf{I}) \text{ and if } \mathbf{Q} \text{ is an orthogonal matrix,}$$
$$\text{then } \mathbf{Q}X \sim N(0, \mathbf{Q}\mathbf{Q}') = N(0, \mathbf{I}). \tag{B.63}$$

c) Densities of Multivariate Normal Distributions

Property B.41: *If* Σ *is nonsingular, then the multivariate normal distribution* $N(m, \Sigma)$ *is absolutely continuous on* $\mathbb{R}^p$ *with density*

$$f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x - m)'\Sigma^{-1}(x - m)\right).$$

PROOF: The multivariate normal distribution $N(m, \Sigma)$ is the probability distribution of $X = \Sigma^{1/2}Z + m$, where the components of Z are independently and identically distributed as $N(0, 1)$. By Property B.24, the density of Z is the product of the marginal densities

$$\ell(z) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} z_i^2\right) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}z'z\right).$$

Because $\Sigma^{1/2}$ is nonsingular, the density of X follows by applying Property B.12. This gives

$$\begin{aligned} f(x) &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(\Sigma^{-1/2}(x-m))'(\Sigma^{-1/2}(x-m))\right) \det(\Sigma^{-1/2}) \\ &= \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x - m)'\Sigma^{-1}(x - m)\right). \end{aligned}$$

□

The preceding result can be generalized to the case where the matrix Σ is singular. In this case the multivariate normal distribution $N(m, \Sigma)$

is absolutely continuous in the affine subspace defined by $m + \mathrm{Im}\Sigma$. Its density is

$$f(x) = \frac{1}{(2\pi)^{\mathrm{rank}\Sigma/2}} \frac{1}{\sqrt{\lambda_1 \ldots \lambda_{\mathrm{rank}\Sigma}}} \exp\left(-\frac{1}{2}(x-m)'\Sigma^-(x-m)\right), \tag{B.64}$$

where $\lambda_1, \ldots, \lambda_{\mathrm{rank}\Sigma}$ denote the nonzero eigenvalues of Σ, and Σ^- is a generalized inverse of Σ. From Property A.25 this expression does not depend on the choice of the generalized inverse Σ^-.

B.3.3 Independence and Conditioning

a) A Necessary and Sufficient Condition for Independence

Consider a Gaussian random vector X, which can be partitioned into K subvectors as $X = (X_1', \ldots, X_K')'$. Similarly, we can partition the mean and the variance covariance matrix of X as $m = (m_1', \ldots, m_K')'$ and

$$VX = \Sigma = \begin{pmatrix} \Sigma_{11} & \ldots & \Sigma_{1K} \\ \vdots & & \vdots \\ \Sigma_{K1} & \ldots & \Sigma_{KK} \end{pmatrix}.$$

Property B.42: *Let $X = (X_1', \ldots, X_K')'$ be a Gaussian random vector. Then the subvectors $X_1, \ldots, X_K$ are independent if and only if $\Sigma_{jk} = 0$, $\forall\, j \neq k$.*

PROOF: Necessity follows from Property B.27. To prove sufficiency, we consider the characteristic function of X. We have

$$\begin{aligned} \phi_X(u) &= E(\exp iu'X) = E\exp\left(i\sum_{k=1}^{K} u_k'X_k\right) \\ &= \exp i\left(u'm - \frac{u'\Sigma u}{2}\right) = \exp i\left(\sum_{k=1}^{K} u_k'm_k - \sum_{k=1}^{K}\frac{u_k'\Sigma_{kk}u_k}{2}\right) \\ &= \prod_{k=1}^{K} \exp\left(iu_k'm_k - \frac{u_k'\Sigma_{kk}u_k}{2}\right) = \prod_{k=1}^{K}\phi_{X_k}(u_k). \end{aligned}$$

The desired result follows. □

Theorem B.4: Cochran Theorem (Part I). *Consider a decomposition of $\mathbb{R}^p$ into the direct sum of K orthogonal subspaces with respect to*

its canonical scalar product. Let $\mathbf{P}_k, k = 1, \ldots, K$, *denote the orthogonal projections on these subspaces. If* X *is a multivariate standard normal Gaussian random vector on* $\mathbb{R}^p$, *i.e.,* $X \sim N(0, \mathbf{I})$, *then the vectors* $X_k = \mathbf{P}_k X$, $k = 1, \ldots, K$, *are independent Gaussian vectors.*

PROOF: First, we note that $(X_1', \ldots, X_K')'$ is a Gaussian random vector since its is obtained through a linear transformation of the Gaussian random vector X. Then, in view of Property B.42 it suffices to show that $\mathrm{Cov}(X_j, X_k) = 0$.

We have

$$\mathrm{Cov}(X_j, X_k) = \mathrm{Cov}(\mathbf{P}_j X, \mathbf{P}_k X) = \mathbf{P}_j V X \mathbf{P}_k' = \mathbf{P}_j \mathbf{P}_k' = \mathbf{P}_j \mathbf{P}_k.$$

But the latter product is zero because the projections are on orthogonal subspaces. □

Theorem B.5: Fisher Theorem (Part I). *Let* $X \sim N(0, \mathbf{I})$. *Then* $\bar{X}_n = (1/n)(X_1 + \ldots + X_n)$ *is independent of* $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

PROOF: We note that $\bar{X}_n$ depends only upon the projection of X on the subspace spanned by the unit vector $e = (1 \ldots 1)'$, while $\hat{\sigma}^2$ depends only on the projection of X on the subspace orthogonal to the preceding subspace (see Property A.11). Then the desired result is an immediate consequence of the Cochran Theorem B.4 and Property B.27-(iv). □

b) Conditioning

Consider a Gaussian random vector, which is partitioned as $(X', Y')'$. Suppose that VX is nonsingular.

Property B.43: *We have:*

(i) $E(Y \mid X) = EY + \mathrm{Cov}(Y, X)(VX)^{-1}(X - EX) = EL(Y \mid X)$.

(ii) $Y - E(Y \mid X)$ *is independent of* X.

PROOF: By definition of the linear regression , the vector $Y - EL(Y \mid X)$ is uncorrelated with X. By property B.42, it is therefore independent of X.

Thus $E(Y - EL(Y \mid X) \mid X) = E(Y - EL(Y \mid X)) = 0$. Then, using the Three-Perpendicular Theorem, we obtain

$$0 = E(Y - EL(Y \mid X) \mid X) = E(Y \mid X) - EL(Y \mid X).$$

□

Hence, in the case of Gaussian random vectors *the conditional expectation and the linear regression coincide.*

Property B.44: *Let $(X',Y')'$ be a Gaussian random vector with a nonsingular variance covariance matrix. Then the conditional probability distribution of Y given $X = x$ is normal with mean $E(Y \mid X = x)$ and variance covariance matrix*

$$\begin{aligned} V(Y \mid X = x) &= VY - VE(Y \mid X) \\ &= VY - \mathrm{Cov}(Y,X)(VX)^{-1}\mathrm{Cov}(X,Y). \end{aligned}$$

In particular, the conditional variance covariance matrix does not depend on x.

PROOF: The conditional density of Y given $X = x$ is $f(y \mid x) = f(x,y)/f_X(x)$, where

$$f_X(x) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{\det VX}} \exp\left(-\frac{1}{2}(x - EX)'(VX)^{-1}(x - EX)\right)$$

and

$$\begin{aligned} f(x,y) &= \frac{1}{(2\pi)^{(p+q)/2}} \frac{1}{\sqrt{\det \Sigma}} \\ &\quad \times \exp\left(-\frac{1}{2}\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} EX \\ EY \end{pmatrix}\right)' \Sigma^{-1} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} EX \\ EY \end{pmatrix}\right)\right). \end{aligned}$$

Now, using the partitioned inverse formulae given in Properties A.4 and A.5, we have

$$\det \Sigma = \det V(X) \det(V(Y) - \mathrm{Cov}(Y,X)(VX)^{-1}\mathrm{Cov}(X,Y))$$

and

$$\begin{aligned} &\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} EX \\ EY \end{pmatrix}\right)' \Sigma^{-1} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} EX \\ EY \end{pmatrix}\right) \\ &= (x - EX)'(VX)^{-1}(x - EX) \\ &\quad + (y - EL(Y \mid X = x))'V(Y \mid X)^{-1}(y - EL(Y \mid X = x)). \end{aligned}$$

The desired result follows. □

B.3.4 Truncated Normal Distributions

a) Deterministic Thresholds

In this subsection we are interested in the probability distribution of the truncated random variable X defined by

$$X = \begin{cases} X^*, \text{if } X^* > 0, \\ 0, \text{ otherwise,} \end{cases} \tag{B.65}$$

where $X^* \sim N(m, \sigma^2)$.

The probability distribution of X is the mixture of a discrete and an absolutely continuous probability distribution on $\mathbb{R}^+$. Specifically, we have

$$\begin{aligned} P(X = 0) &= P(X^* \leq 0) = \Phi(-m/\sigma) \\ P(a < X < b) &= \int_a^b \frac{1}{\sigma}\phi\left(\frac{x-m}{\sigma}\right) dx, \text{ if } \sigma > 0. \end{aligned} \tag{B.66}$$

The moments and the conditional moments of this truncated variable can be computed by using the next formulae.

Property B.45: *Let $X = X^* \mathbb{1}_{X^*>0}$, where $X^* \sim N(m, \sigma^2)$. Then we have:*

(i) $EX = \sigma\phi(m/\sigma) + m\Phi(m/\sigma)$,

(ii) $E(X \mid X > 0) = m + \sigma\frac{\phi(m/\sigma)}{\Phi(m/\sigma)}$,

(iii) $VX = \sigma^2\left(\Phi(\frac{m}{\sigma}) + \frac{m}{\sigma}W(\frac{m}{\sigma}) - W^2(\frac{m}{\sigma})\right)$, *where* $W(X) = \int_{-\infty}^{x}\Phi(t)dt = x\Phi(x) + \phi(x)$,

(iv) $V(X \mid X > 0) = \sigma^2\left(1 - \frac{m}{\sigma}\lambda(\frac{m}{\sigma}) - \lambda^2(\frac{m}{\sigma})\right)$, *where* $\lambda(x) = \phi(x)/\Phi(x)$ *is called the Mill's ratio.*

PROOF: We prove (i) only. The other statements can be proved similarly. We have

$$\begin{aligned} EX &= \int_0^\infty x\frac{1}{\sigma}\phi\left(\frac{x-m}{\sigma}\right) dx \\ &= \int_0^\infty \frac{x-m}{\sigma}\phi\left(\frac{x-m}{\sigma}\right) dx + m\int_0^\infty \frac{1}{\sigma}\phi\left(\frac{x-m}{\sigma}\right) dx \end{aligned}$$

$$
\begin{aligned}
&= \sigma \int_{-m/\sigma}^{\infty} t\phi(t)dt + m \int_{-m/\sigma}^{\infty} \phi(t)dt \\
&= \sigma[-\phi(t)]_{-m/\sigma}^{\infty} + m[\Phi(t)]_{-m/\sigma}^{\infty} \\
&= \sigma\phi\left(\frac{m}{\sigma}\right) + m\Phi\left(\frac{m}{\sigma}\right) = \sigma W\left(\frac{m}{\sigma}\right).
\end{aligned}
$$

□

The Mill's ratio plays an important role in the study of truncated normal distributions. It is useful to know how $\lambda(x)$ varies as a function of x. We have

$$
\frac{d\lambda(x)}{dx} = -\lambda(x)(x + \lambda(x)) \equiv -\frac{\lambda(x)W(x)}{\Phi(x)}.
$$

Note that $W(x)$ is strictly positive since it is the integral of a strictly positive function. Thus, because the three terms of the preceding equality are strictly positive, then the function $\lambda(\cdot)$ is strictly decreasing. Its behavior when $x \to \pm\infty$ follows from Property B.33-(v)(vi). Its graph is illustrated in Figure B.3.

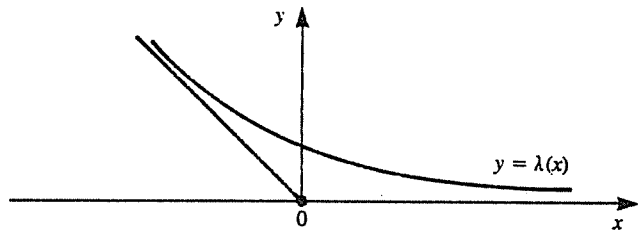


Figure B.3: Mill's Ratio

b) Random Thresholds

The preceding results can be generalized to the case where a random threshold truncates the normal random variable. Specifically, we now consider the random variable X defined by

$$
X = \begin{cases} X^*, \text{ if } Y^* > 0 \\ 0, \text{ otherwise,} \end{cases} \tag{B.67}
$$

where

$$
\begin{pmatrix} X^* \\ Y^* \end{pmatrix} \sim N\left(\begin{pmatrix} m_x \\ m_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right).
$$

We note that this does generalize the preceding case since the threshold, which was zero previously, is now given by $Z^* = X^* - Y^*$. For we have

$$X = X^* \mathbb{I}_{Y^*>0} = X^* \mathbb{I}_{X^*>Z^*}.$$

Property B.46: *The unconditional and conditional expectations of X, as defined by equation (B.66), are respectively*

$$EX = m_x \Phi(m_y/\sigma_y) + \rho\sigma_x \phi(m_y/\sigma_y)$$

and

$$E(X \mid Y^* > 0) = m_x + \rho\sigma_x \lambda(m_y/\sigma_y).$$

PROOF: We prove the first equality only. It is useful to introduce the conditional expectation

$$E(X^* \mid Y^*) = m_x + \frac{\rho\sigma_x\sigma_y}{\sigma_y^2}(Y^* - m_y).$$

Then we can write

$$\begin{aligned} E(X) &= P(Y^* > 0)E(X \mid Y^* > 0) \\ &= P(Y^* > 0)E(X^* \mid Y^* > 0) \\ &= P(Y^* > 0) \\ &\quad \times (E(E(X^*|Y^*)|Y^* > 0) + E(X^* - E(X^*|Y^*)|Y^* > 0)). \end{aligned}$$

In view of Property B.43-(ii) we know that $X^* - E(X^* \mid Y^*)$ and Y^* are independent. Thus we have

$$\begin{aligned} E(X^* - E(X^* \mid Y^*) \mid Y^* > 0) &= E(X^* - E(X^* \mid Y^*)) \\ &= EX^* - EE(X^* \mid Y^*) = 0. \end{aligned}$$

Hence

$$\begin{aligned} EX &= P(Y^* > 0)E(E(X^* \mid Y^*) \mid Y^* > 0) \\ &= P(Y^* > 0)\left(m_x - \frac{\rho\sigma_x\sigma_y}{\sigma_y^2}m_y + \rho\frac{\sigma_x}{\sigma_y}E(Y^* \mid Y^* > 0)\right) \\ &= \Phi(m_y/\sigma_y)\left(m_x + \rho\frac{\sigma_x}{\sigma_y}\sigma_y\frac{\phi(m_y/\sigma_y)}{\Phi(m_y/\sigma_y)}\right) \\ &= m_x\Phi(m_y/\sigma_y) + \rho\sigma_x\phi(m_y/\sigma_y). \end{aligned}$$

□

c) Minimum of Two Normal Random Variables

We use the notation of the preceding subsection. We now consider the probability distribution of $Z = \min(X^*, Y^*)$. To simplify, we assume that the random variables X^* and Y^* are independent.

Property B.47: *Let X^* and Y^* be two independent random variables distributed as $N(m_x, \sigma_x^2)$ and $N(m_y, \sigma_y^2)$, respectively. The random variable $Z = \min(X^*, Y^*)$ has an absolutely continuous probability distribution on $\mathbb{R}$ with density*

$$\begin{aligned} f(z) &= \frac{1}{\sigma_x}\phi\left(\frac{z-m_x}{\sigma_x}\right)\left(1-\Phi\left(\frac{z-m_y}{\sigma_y}\right)\right) \\ &\quad +\frac{1}{\sigma_y}\phi\left(\frac{z-m_y}{\sigma_y}\right)\left(1-\Phi\left(\frac{z-m_x}{\sigma_x}\right)\right). \end{aligned}$$

PROOF: Let F_X and F_Y denote the cumulative distribution function of X^* and Y^*, respectively. Let f_X and f_Y denote their densities. We have

$$\begin{aligned} P(Z > z) &= P(\min(X^*, Y^*) > z) = P(X^* > z \text{ and } Y^* > z) \\ &= P(X^* > z)P(Y^* > z) = (1 - F_X(z))(1 - F_Y(z)). \end{aligned}$$

Because the density of z is $f(z) = -dP(Z > z)/dz$, then we obtain $f(z) = f_X(z)(1 - F_Y(z)) + f_Y(z)(1 - F_X(z))$. It now suffices to use the expressions of f_X, f_Y, F_X, and F_Y to obtain the desired result. □

Property B.48: *We have*

$$EZ = m_x - \sqrt{\sigma_x^2 + \sigma_y^2}\ \phi\left(\frac{m_x - m_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right) - (m_x - m_y)\ \Phi\left(\frac{m_x - m_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right).$$

PROOF: We have

$$\begin{aligned} EZ &= E\min(X^*, Y^*) = E(X^* \mathbb{1}_{X^* < Y^*}) + E(Y^* \mathbb{1}_{X^* > Y^*}) \\ &= EX^* - E((X^* - Y^*)\mathbb{1}_{X^* - Y^* > 0}). \end{aligned}$$

Because $X^* - Y^* \sim N(m_x - m_y, \sigma_x^2 + \sigma_y^2)$, then we can apply Property B.45-(i). This establishes the desired result. □

B.4 Distributions Derived from the Normal Distribution

B.4.1 Chi-Square Distributions

Definition B.21: *A chi-square distribution with K degrees of freedom is the probability distribution of the random variable $Y = X_1^2 + \ldots + X_K^2 = \|X\|^2$, where the random variables X_k are independently distributed with respective distributions $N(m_k, 1)$. When $m_k = 0$, $\forall\ k$, the chi-square distribution is said to be central. It is noncentral otherwise.*

Property B.49: *The probability distribution of the random variable $Y = X_1^2 + \ldots + X_K^2$, where X_k are independently distributed $N(m_k, 1)$, depends only on K and $\lambda = \sum_{k=1}^{K} m_k^2 = \|m\|^2$. Hence this probability distribution can be denoted $\chi^2(K, \lambda)$, where λ is called the noncentrality parameter. A central chi-square distribution is denoted $\chi^2(K) = \chi^2(K, 0)$.*

PROOF: Consider the transformation associated with some orthogonal matrix $\mathbf{Q}$. Define $Z = \mathbf{Q}X$. Then, by (B.63) we have $Z \sim N(\mathbf{Q}m, \mathbf{I})$. Moreover, $Y = \|X\|^2 = Z_1^2 + \ldots + Z_K^2$ because the Euclidian length remains unchanged under an orthogonal transformation. Then choose the matrix $\mathbf{Q}$ so that its first row is equal to $m'/\|m\|$. This is possible as soon as $m \neq 0$. The other rows of $\mathbf{Q}$ are therefore orthogonal to m. Hence

$$\mathbf{Q}m = \begin{pmatrix} \|m\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Now the probability distribution of Y is equal to the probability distribution of $Z_1^2 + \ldots + Z_K^2$, where the Z_k's are independent random variables. Moreover, we have $Z_1 \sim N(\|m\|, 1)$ and $Z_k \sim N(0, 1)$ for $k \geq 2$. Hence this distribution depends on m through $\|m\|$ only. □

From Definition B.21 and Property B.49 it follows at once that the family of chi-square distributions is closed under the addition of independent random variables.

Property B.50: *Let Y_1 and Y_2 be two independent random variables distributed as $\chi^2(K_1, \lambda_1)$ and $\chi^2(K_2, \lambda_2)$, respectively. Then their sum $Y = Y_1 + Y_2$ is distributed as $\chi^2(K_1 + K_2, \lambda_1 + \lambda_2)$.*

Property B.51: *Let $Y \sim \chi^2(K, \lambda)$. Then $EY = K + \lambda$ and $VY = 2(K + 2\lambda)$. In particular, if $Y \sim \chi^2(K)$, then $EY = K$ and $VY = 2K$.*

PROOF: We use the orthogonal transformation $\mathbf{Q}$ of the proof of Property B.49. We have $EY = EZ_1^2 + \sum_{k=2}^{K} E(Z_k^2)$. Let Z denote a standard normal random variable. Then we have

$$EY = E(Z + \sqrt{\lambda})^2 + \sum_{k=2}^{K} E(Z^2) = 1 + (\sqrt{\lambda})^2 + (K-1) = K + \lambda.$$

Similarly, we have $VY = V(Z_1^2) + \sum_{k=2}^{K} V(Z_k^2)$. Hence

$$\begin{aligned} VY &= E(Z + \sqrt{\lambda})^4 - (E(Z + \sqrt{\lambda})^2)^2 + (K-1)(E(Z^4) - E(Z^2)^2) \\ &= EZ^4 + 6\lambda EZ^2 + \lambda^2 - (1+\lambda)^2 + (K-1)(EZ^4 - (EZ^2)^2) \\ &= 2 + 4\lambda + 2(K-1) = 2(K + 2\lambda), \end{aligned}$$

where we have used equation (B.58). □

The form of the density of a chi-square distribution can be determined by noting that the family of chi-square distributions is related to the family of gamma distributions. See Table B.2. The next property is given without proof.

Property B.52: *The random variable Y is distributed as $\chi^2(K)$ if and only if $Z = Y/2$ is distributed as a $\gamma(K/2)$. In particular, the density of a (central) chi-square distribution with K degrees of freedom is*

$$f(y) = \mathbb{1}_{\mathbb{R}^+}(y) \frac{1}{\Gamma(K/2)} \frac{y^{K/2-1}}{2^{K/2}} \exp -(y/2).$$

A similar property holds for a noncentral chi-square distribution Specifically, the density of the $\chi^2(K, \lambda)$ distribution is

$$f(y) = \mathbb{1}_{\mathbb{R}^+}(y) \exp\left(-\frac{\lambda}{2}\right) \exp\left(-\frac{y}{2}\right) \frac{y^{K/2-1}}{2^{K/2}} \sum_{i=0}^{\infty} \frac{y^i \lambda^i}{i!\, 2^{2i}} \frac{1}{\Gamma(i + K/2)}. \tag{B.68}$$

The latter expression follows from the next property, which allows us to interpret a noncentral chi-square distribution as a central chi-square distribution whose number of degrees of freedom is random.

Property B.53: *Let $X_1, \ldots, X_n, \ldots$ be independent random variables distributed as $N(0,1)$. Let Z be a random variable independent of $X_1, \ldots, X_n, \ldots$ and distributed as a Poisson $\mathcal{P}(\lambda/2)$. Then the random variable $Y = X_1^2 + \ldots + X_{K+2Z}^2$ follows a $\chi^2(K, \lambda)$.*

Note that the number of elements in the sum is random and equal to $K + 2Z$. The proof of this property relies upon the following lemma, which gives the characteristic function of the $\chi^2(K, \lambda)$ distribution.

Lemma B.1: *If* $Y \sim \chi^2(K, \lambda)$, *then*

$$\phi_Y(t) = \frac{1}{(1-2it)^{K/2}} \exp(-\lambda/2) \exp\left(\frac{\lambda}{2(1-2it)}\right).$$

PROOF OF THE LEMMA: Let $U_1, \ldots, U_K$ denote K independent random variables distributed as $N(0,1)$. We have

$$\begin{aligned}
\phi_Y(t) &= E(\exp itY) = E\left(\exp it \sum_{k=1}^{K} (U_k + m_k)^2\right) \\
&= \prod_{k=1}^{K} E(\exp it(U_k + m_k)^2) \\
&= \prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{u^2}{2} + it(u + m_k)^2\right) du\right) \\
&= \prod_{k=1}^{K} \frac{1}{\sqrt{1-2it}} \exp\left(\frac{itm_k^2}{1-2it}\right) \\
&= \frac{1}{(1-2it)^{K/2}} \exp\left(\frac{\lambda it}{1-2it}\right) \\
&= \frac{1}{(1-2it)^{K/2}} \exp(-\lambda/2) \exp \frac{\lambda}{2(1-2it)},
\end{aligned}$$

where the second equality follows from the independence of the U_k's. □

PROOF OF PROPERTY B.53: The characteristic function of Y can be written as

$$\phi_Y(t) = \exp(-\lambda/2) \sum_{j=0}^{\infty} (\lambda/2)^j \frac{1}{j!} \frac{1}{(1-2it)^{(K+2j)/2}}.$$

Now, we know that $1/((1-2it)^{(K+2j)/2})$ is the characteristic function of the $\chi^2(K+2j)$ distribution. We also note that the corresponding weight $(1/j!)\exp(-\lambda/2)(\lambda/2)^j$ is the elementary probability of the Poisson distribution $\mathcal{P}(\lambda/2)$.

Thus it suffices to show that the characteristic function of the sum is equal to this quantity. We have

$$\begin{aligned} & E(\exp it(X_1^2 + \ldots + X_{K+2Z}^2)) \\ &= EE(\exp it(X_1^2 + \ldots + X_{K+2z}^2) \mid Z = z) \\ &= \sum_{j=0}^{\infty} P(Z = j) E(\exp it(X_1^2 + \ldots + X_{K+2j}^2)). \end{aligned}$$

The desired result follows from Theorem B.1. □

B.4.2 Quadratic Forms of Gaussian Vectors

Theorem B.6: Cochran Theorem (Part II). *Let X be a random vector of size p distributed as $N(m, \mathbf{I})$. Consider a decomposition of $\mathbb{R}^p$ into the direct sum of J orthogonal subspaces. For every $j = 1, \ldots, J$, let $\mathbf{P}_j$ denote the orthogonal projection on the j-th subspace. Let $K_j = \text{rank}\mathbf{P}_j$ be its dimension. Then the random variables $\|\mathbf{P}_j X\|^2$, $j = 1, \ldots, J$, are independent. Moreover, $\|\mathbf{P}_j X\|^2 \sim \chi^2(K_j, \|\mathbf{P}_j m\|^2)$.*

PROOF: The independence follows immediately from Part I of Cochran Theorem. See Theorem B.4. It remains to derive the probability distribution of $\|P_1 X\|^2$ (say).

To do so, we begin by applying an orthogonal transformation $\mathbf{Q}$ that allows us to use an orthonormal basis composed of vectors belonging to the J subspaces. We have $X = \mathbf{Q}Z$, where Z is the vector of new coordinates. Moreover, $Z \sim N(\mathbf{Q}'m, \mathbf{I})$ and we have

$$\mathbf{Q}'\mathbf{P}_1 X = \mathbf{Q}'\mathbf{P}_1\mathbf{Q}\mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_{K_1} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

provided the first K_1 basis vectors belong to the first subspace.

It follows that $\|\mathbf{P}_1 X\|^2 = \|\mathbf{Q}'\mathbf{P}_1 X\|^2 = Z_1^2 + \ldots + Z_{K_1}^2$. Hence, because $(EZ_1)^2 + \ldots + (EZ_{K_1})^2 = \|E(\mathbf{Q}'\mathbf{P}_1 X)\|^2 = \|\mathbf{Q}'\mathbf{P}_1 m\|^2 = \|\mathbf{P}_1 m\|^2$, we obtain $\|\mathbf{P}_1 X\|^2 \sim \chi^2(K_1, \|\mathbf{P}_1 m\|^2)$. □

As a special case, consider the subspace spanned by the vector $e = (1, \ldots, 1)'$ and its orthogonal subspace. Then we obtain the following result.

Theorem B.7: Fisher Theorem (Part II). *Let X_i, $i = 1, \ldots, n$, be independent random variables distributed as $N(m_i, \sigma^2)$. The sample mean $X_n = (1/n)\sum_{i=1}^n \bar{X}_i$ and the sample variance $S_n^2 = 1/(n-1)\sum_{i=1}^n (\bar{X}_i - X_n)^2$ are independent. Moreover, $\bar{X}_n$ is distributed as $N(\bar{m}_n, \sigma^2/n)$, where $\bar{m}_n = (1/n)\sum_{i=1}^n m_i$, while $(n-1)S_n^2/\sigma^2$ is distributed as $\chi^2(n-1, \lambda)$, where $\lambda = \sum_{i=1}^n (m_i - \bar{m}_n)^2/\sigma^2$.*

Moreover, the Cochran Theorem B.6 implies that, if X is distributed as $N(m, \mathbf{I})$ and if $\mathbf{P}$ is an orthogonal projection, then $\|\mathbf{P}X\|^2 = X'\mathbf{P}X \sim \chi^2(\text{rank}\mathbf{P}, \|\mathbf{P}m\|^2)$. The converse also holds.

Property B.54: *Let X be a Gaussian random vector distributed as $N(m, \mathbf{I})$. Let $\mathbf{P}$ be a symmetric matrix. If $X'\mathbf{P}X$ is distributed as a chi-square, then $\mathbf{P}^2 = \mathbf{P}$, i.e., $\mathbf{P}$ is the matrix of an orthogonal projection.*

PROOF: First, we note that we can always use a basis of eigenvectors of the matrix $\mathbf{P}$ after a suitable change of orthonormal basis, i.e., after applying a suitable orthogonal transformation. Thus we can assume that the matrix $\mathbf{P}$ is diagonal. Let λ_j, $j = 1, \ldots, p$, be its eigenvalues. Then, with respect to this basis the quadratic form $X'\mathbf{P}X$ is of the form $\sum_{j=1}^p \lambda_j (U_j + \mu_j)^2$, where the random variables U_j are independent and identically distributed as $N(0,1)$, and where μ_j are some scalars.

Next, an argument similar to the one used in Lemma B.1 allows us to obtain the characteristic function of the above quadratic form. Namely, we have

$$E \exp\left(it \sum_{j=1}^p \lambda_j (U_j + \mu_j)^2 \right)$$
$$= \prod_{j=1}^p \frac{1}{\sqrt{1 - 2it\lambda_j}} \exp(-\frac{\mu}{2}) \left(\frac{1}{2} \sum_{j=1}^p \frac{\mu_j^2}{1 - 2it\lambda_j} \right),$$

where $\mu = \sum_{j=1}^p \mu_j^2$. Thus we need to consider the case where this characteristic function is identical to the characteristic function of a $\chi^2(K^*, \mu^*)$, i.e., to

$$\phi(t) = \frac{1}{(1-2it)^{K^*/2}} \exp\left(-\frac{\mu^*}{2}\right) \exp\left(\frac{1}{2}\frac{\mu^*}{1-2it}\right),$$

where we have used Lemma B.1.

It is easy to see that such an equality of characteristic functions holds only if the eigenvalues λ_j are equal to either zero or one. In this case K^*

is equal to the number of nonzero eigenvalues. Because all its eigenvalues are equal to zero or one, the matrix $\mathbf{P}$ is therefore idempotent. □

The preceding property can be easily generalized to a Gaussian random vector whose variance–covariance matrix is not the identity matrix. For it suffices to apply an appropriate linear transformation so as to use Property B.54.

Corollary B.3: *Let Y be distributed as $N(m, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is nonsingular. Let $\mathbf{\Pi}$ be a symmetric matrix. Then $Y'\mathbf{\Pi}Y$ is distributed as a chi-square if and only if $\mathbf{\Pi\Sigma\Pi} = \mathbf{\Pi}$, in which case the chi-square distribution is $\chi^2(\text{rank}\mathbf{\Pi}, m'\mathbf{\Pi}m)$.*

When $\mathbf{\Sigma}$ is singular, the latter property is generalized as follows.

Corollary B.4: *Let Y be a Gaussian random vector distributed as $N(m, \mathbf{\Sigma})$. Then the quadratic form $Y'\mathbf{\Pi}Y$ is distributed as a chi-square if and only if:*

(i) $\mathbf{\Sigma\Pi\Sigma\Pi\Sigma} = \mathbf{\Sigma\Pi\Sigma}$,

(ii) $m'\mathbf{\Pi\Sigma\Pi}m = m'\mathbf{\Pi}m$, *and*

(iii) $\mathbf{\Sigma\Pi\Sigma\Pi}m = \mathbf{\Sigma\Pi}m$.

If this is the case, then the number of degrees of freedom of the chi-square distribution is equal to $\text{rank}(\mathbf{\Sigma\Pi\Sigma})$, *and its noncentrality parameter is equal to* $m'\mathbf{\Pi}m$.

Note that the preceding conditions are satisfied when m is equal to zero and $\mathbf{\Pi}$ is a symmetric generalized inverse $\mathbf{\Sigma}^-$ of $\mathbf{\Sigma}$. This gives the next corollary.

Corollary B.5: *If Y is distributed as $N(m, \mathbf{\Sigma})$, then $(Y-m)'\mathbf{\Sigma}^-(Y-m)$ is distributed as $\chi^2(\text{rank}\mathbf{\Sigma})$.*

B.4.3 Student Distributions

Definition B.22: *The Student distribution with K degrees of freedom is the probability distribution of the random variable $Z = X/\sqrt{Y/K}$, where X and Y are independent random variables distributed as $N(m, 1)$ and $\chi^2(K)$, respectively. The parameter m is called the noncentrality parameter. The central Student distribution corresponds to $m = 0$ and is denoted $t(K)$.*

Property B.55: *The Student distribution with K degrees of freedom and noncentrality parameter m is absolutely continuous on $\mathbb{R}$ with density*

$$\begin{aligned} f(z) &= \frac{K^{K/2}}{\Gamma(K/2)\Gamma(1/2)} \frac{\exp(-m^2/2)}{(K+z^2)^{(K+1)/2}} \\ &\quad \times \sum_{i=0}^{\infty} \Gamma\left(\frac{K+i+1}{2}\right) \frac{m^i}{i!} \left(\frac{2z^2}{K+z^2}\right)^{i/2} (\text{sign}z)^i. \end{aligned}$$

In particular, the density of the central Student distribution $t(K)$ is

$$f(z) = \frac{\Gamma\left(\frac{K+1}{2}\right)}{\Gamma\left(\frac{K}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{1}{\sqrt{K}} \frac{1}{\left(1+\frac{z^2}{K}\right)^{(K+1)/2}}.$$

PROOF: The density of the probability distribution of the pair (X, Y) is

$$f(x,y) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x-m)^2}{2} \frac{1}{\Gamma(K/2)} \frac{1}{2^{K/2}} \exp\left(-\frac{y}{2}\right) y^{K/2-1}.$$

Consider now the polar coordinates

$$x = r\sin\theta, \quad y = r^2\cos^2\theta, \quad dxdy = 2r^2\cos\theta dr d\theta.$$

The joint density of (r, θ) is

$$\begin{aligned} f(r,\theta) &= \frac{1}{\Gamma(1/2)} \frac{1}{\Gamma(K/2)} \frac{2}{2^{(K+1)/2}} \exp(-m^2/2) \exp(-r^2/2) \\ &\quad \times \exp(mr\sin\theta) r^K (\cos\theta)^{K-1}. \end{aligned}$$

Then, using the series expansion

$$\exp(mr\sin\theta) = \sum_{i=0}^{\infty} \frac{m^i}{i!} r^i (\sin\theta)^i,$$

we obtain

$$\begin{aligned} f(r,\theta) &= \frac{1}{\Gamma(1/2)} \frac{1}{\Gamma(K/2)} \frac{1}{2^{(K-1)/2}} \exp(-m^2/2) \exp(-r^2/2) \\ &\quad \times (\cos\theta)^{K-1} \sum_{i=0}^{\infty} \frac{m^i}{i!} r^{K+i} (\sin\theta)^i. \end{aligned}$$

Because

$$\int_0^\infty \exp(-r^2/2) r^{K+i} dr = 2^{(K+i-1)/2} \Gamma((K+i+1)/2),$$

then integrating the joint density $f(r, \theta)$ with respect to r gives the marginal density of θ

$$\begin{aligned} f(\theta) &= \frac{1}{\Gamma(1/2)} \frac{1}{\Gamma(K/2)} \exp(-m^2/2)(\cos\theta)^{K-1} \\ &\quad \times \sum_{i=0}^{\infty} \Gamma\left(\frac{K+i+1}{2}\right) \frac{m^i}{i!} 2^{i/2} (\sin\theta)^i. \end{aligned}$$

Lastly, because $z = \sqrt{K} \tan\theta$, the density of z is obtained after a change of variable. This gives

$$\begin{aligned} f(z) &= \frac{1}{\Gamma(1/2)} \frac{1}{\Gamma(K/2)} \exp(-m^2/2) K^{K/2} \frac{1}{(K+z^2)^{(K+1)/2}} \\ &\quad \times \sum_{i=0}^{\infty} \Gamma\left(\frac{K+i+1}{2}\right) \frac{m^i}{i!} \left(\frac{2z^2}{K+z^2}\right)^{i/2} (\text{sign} z)^i. \end{aligned}$$

□

The central Student distribution has a density that is symmetric about zero. Its moment of order p exists only if $p < K$. It is given by

$$EZ^p = 0 \quad \text{if } p \text{ is odd with } p < K$$

and

$$EZ^p = K^{p/2} \frac{\Gamma\left(\frac{p+1}{2}\right) \Gamma\left(\frac{K-p}{2}\right)}{\Gamma(1/2)\Gamma(K/2)} \quad \text{if } p \text{ is even with } p < K. \tag{B.69}$$

In particular, the variance of a Student $t(K)$ for $K > 2$ is

$$VZ = EZ^2 = K \frac{\Gamma(3/2)\Gamma(K/2-1)}{\Gamma(1/2)\Gamma(K/2)} = \frac{K}{K-2}.$$

B.4.4 Fisher Distributions

Definition B.23: *The Fisher distribution with K_1 and K_2 degrees of freedom is the probability distribution of the random variable $Z = (Y_1/K_1)/(Y_2/K_2)$, where Y_1 and Y_2 are independent random variables distributed as $\chi^2(K_1)$ and $\chi^2(K_2)$, respectively. This distribution is denoted $F(K_1, K_2)$.*

Example B.1: Let $X_1, \ldots, X_n$ be n independent random variables distributed as $N(0, \sigma^2)$. Let $\mathbb{R}^n = E_1 \oplus E_2$ be a decomposition of $\mathbb{R}^n$ into two orthogonal subspaces. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be the orthogonal projection matrices on E_1 and E_2, respectively. From the Cochran Theorem B.6, the random variables $\|\mathbf{P}_1 X\|^2/\sigma^2$ and $\|\mathbf{P}_2 X\|^2/\sigma^2$, where $\mathbf{X} = (X_1 \ldots X_n)'$, are independent and are distributed as $\chi^2(\dim E_1)$ and $\chi^2(\dim E_2) = \chi^2(n - \dim E_1)$, respectively. Thus the random variable

$$Z = \frac{\|\mathbf{P}_1 X\|^2/\dim E_1}{\|\mathbf{P}_2 X\|^2/\dim E_2}$$

follows an $F(\dim E_1, \dim E_2)$ distribution.

Remark B.1: If $\tilde{Z}$ follows a central Student distribution $t(K)$, then its square $Z = \tilde{Z}^2$ follows an $F(1, K)$ distribution.

As for the Student distribution, successive change of variables allows us to derive the density of the Fisher distribution.

Property B.56: *The Fisher distribution $F(K_1, K_2)$ is absolutely continuous on $\mathbb{R}^+$ with density*

$$f(z) = (K_1)^{K_1/2}(K_2)^{K_2/2} \frac{\Gamma\left(\frac{K_1+K_2}{2}\right) z^{K_1/2-1}}{\Gamma\left(\frac{K_1}{2}\right)\Gamma\left(\frac{K_2}{2}\right)(K_2 + K_1 z)^{(K_1+K_2)/2}}.$$

The graph of this density is given in Figure B.4.

The moments of the Fisher distribution $F(K_1, K_2)$ exist only up to orders strictly smaller than $K_2/2$. They are given by

$$\text{if } Z \sim F(K_1, K_2) : EZ^p = \left(\frac{K_2}{K_1}\right)^p \frac{\Gamma\left(\frac{K_1}{2}+p\right)\Gamma\left(\frac{K_2}{2}-p\right)}{\Gamma\left(\frac{K_1}{2}\right)\Gamma\left(\frac{K_2}{2}\right)}. \quad \text{(B.70)}$$

In particular, its expectation is $EZ = K_2/(K_2 - 2)$ and its variance is

$$VZ = \frac{2K_2^2(K_1 + K_2 - 2)}{K_1(K_2 - 2)^2(K_2 - 4)}.$$

More generally, we can define the noncentral Fisher distribution.

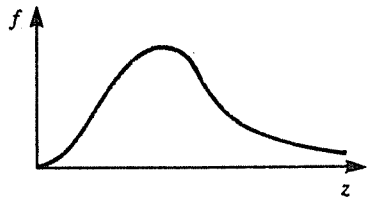


Figure B.4: Density of a Fisher Distribution

Definition B.24: *The noncentral Fisher distribution with K_1 and K_2 degrees of freedom and noncentrality parameters λ_1 and λ_2 is defined as the probability distribution of*

$$Z = \frac{Y_1/K_1}{Y_2/K_2},$$

where Y_1 and Y_2 are independent random variables that are distributed as $\chi^2(K_1, \lambda_1)$ and $\chi^2(K_2, \lambda_2)$, respectively. This Fisher distribution is denoted $F(K_1, K_2; \lambda_1, \lambda_2)$.

In statistical applications the second noncentrality parameter is typically equal to zero. Provided $K_2 > 4$, the first two moments of the $F(K_1, K_2; \lambda_1, 0)$ distribution are

$$\begin{aligned} EZ &= \frac{K_2(K_1 + \lambda_1)}{K_1(K_2 - 2)}, \\ VZ &= 2\left(\frac{K_2}{K_1}\right)^2 \frac{(K_1 + \lambda_1)^2 + (K_1 + 2\lambda_1)(K_2 - 2)}{(K_2 - 2)^2(K_2 - 4)}. \end{aligned}$$

B.5 Stochastic Processes

B.5.1 Definition

Definition B.24: *A stochastic process is a sequence $\{X_n, n \in I\}$ of random variables defined on a same probability space $(\Omega, \mathcal{A}, P)$ with values in $\mathbb{R}^p$.*

The set of indices I is typically $\mathbb{N}$, $\mathbb{N}^*$, or $\mathbb{Z}$. Thus to each state of Nature $\omega \in \Omega$ corresponds an element $\{X_n(\omega), n \in I\}$ of $\mathbb{R}^{pI}$. Such an element is called a path of the stochastic process.

A stochastic process has a probability distribution. It can be shown that this distribution is fully characterized by the knowledge of the probability distributions of all finite subfamilies $\{X_{n_1}, \ldots, X_{n_k}\}$ of the sequence $\{X_n, n \in I\}$. Such a simple characterization allows us to introduce some special classes of stochastic processes.

Definition B.26:

(i) A stochastic process is a sequence of independent random vectors if every finite subfamily $\{X_{n_1}, \ldots, X_{n_k}\}$ is composed of independent random vectors.

(ii) If in addition the probability distribution of X_n is independent of the index n, then the process is composed of independent and identically distributed random vectors. Such a sequence is called a random sample.

(iii) If in addition the common probability distribution is centered with a finite second-order moment, then the process is called an (independent) white noise.

Definition B.27: *A stochastic process is Gaussian if every of its finite subfamilies is distributed as a multivariate normal distribution.*

Definition B.27 allows us to generalize the concept of normal distribution to sequences. Because the probability distribution of a stochastic process is fully characterized by the probability distributions of its finite subfamilies, and because the latter distributions (when they are normal) are fully characterized by their means and variance covariance matrices, it follows that the probability distribution of a Gaussian process is known as soon as we know the sequence of means $\{m_n = EX_n, n \in I\}$ and all the variance covariance matrices $\Gamma_{n,m} = \text{Cov}(X_n, X_m), n, m \in I$.

Definition B.28: *A stochastic process is stationary if the probability distribution of the process is invariant under translation of the index, i.e., if the probability distribution of $(X_{n_1}, \ldots, X_{n_k})$ is identical to that of $(X_{n_1+h}, \ldots, X_{n_k+h})$, $\forall h, k, n_1, \ldots, n_k$.*

As we shall see below, stationary processes are important for the derivation of asymptotic results. Some simple examples of stationary processes are (i) sequences of independent and identically distributed random vectors and (ii) Gaussian processes for which $m_n = m$ (independent of n) and $\Gamma_{n+h,m+h} = \Gamma_{n,m}$, $\forall\ n, m, h$.

B.5.2 Types of Convergence of Random Vectors

In this subsection we consider a sequence of random vectors $\{X_n, n \in \mathbb{N}\}$ and we are interested in its convergence properties as n increases to $+\infty$.

There exist various possible definitions of convergence. They are more or less restrictive and more or less easy to verify and use.

To simplify, we assume that the X_n's are square integrable. We let $\|\cdot\|$ denote the usual Euclidean norm on $\mathbb{R}^p$.

Definition B.29: *The sequence X_n converges in probability (or weakly) to the random vector X if*

$$\forall\, \epsilon > 0,\ P(\|X_n - X\| > \epsilon) \to 0,$$

as n increases to ∞. This type of convergence is denoted $X_n \xrightarrow{P} X$.

Definition B.30: *The sequence X_n converges almost surely (or strongly) to the random vector X if the probability of the event that X_n converges simply to X is equal to one, or equivalently, if*

$$\forall \epsilon > 0,\ P(\sup_{n \geq N} \|X_n - X\| > \epsilon) \to 0.$$

as N increases to ∞. This type of convergence is denoted $X_n \xrightarrow{a.s.} X$.

It is clear that the second type of convergence is more stringent than the first one. For we have

$$X_n \xrightarrow{a.s.} X \ \Rightarrow\ X_n \xrightarrow{P} X, \tag{B.71}$$

i.e., convergence almost surely implies convergence in probability.

Another type of convergence, which is easier to apply, is based on the second-order moments.

Definition B.31: *The sequence X_n converges in quadratic mean (or in L^2) to the random vector X if*

$$E\|X_n - X\|^2 \to 0,$$

as n increases to ∞. This type of convergence is denoted $X_n \xrightarrow{q.m.} X$.

It is easy to see that the preceding condition is equivalent to the convergence of the matrices $E(X_n - X)(X_n - X)'$ to the zero matrix. Moreover, it follows from the usual properties of Hilbert spaces that a sufficient condition for convergence in quadratic mean is that the series $X_{n+1} - X_n$ converges in norm, i.e.

$$\left(\sum_{n=1}^{\infty} \sqrt{E\|X_{n+1} - X_n\|^2} < \infty\right) \Rightarrow X_n \xrightarrow{q.m.} X. \tag{B.72}$$

Convergence in quadratic mean can be easily related to convergence in probability using the following well-known inequality.

Property B.57: Bienaymé–Tchebychev Inequality. *We have*

$$\forall\, \epsilon > 0,\ P(\|X_n - X\| > \epsilon) \leq \frac{E\|X_n - X\|^2}{\epsilon^2}.$$

Thus, if $E\|X_n - X\|^2$ converges to zero, then $P(\|X_n - X\| > \epsilon)$ also converges to zero. Therefore we have the following implication

$$X_n \xrightarrow{q.m.} X \ \Rightarrow\ X_n \xrightarrow{P} X,$$

i.e., convergence in quadratic mean implies convergence in probability.

Lastly, we note that convergence in quadratic mean is quite easy to verify when the limit is nonrandom.

Property B.58: *We have $X_n \xrightarrow{q.m.} a \in \mathbb{R}^p$ if and only if*

(i) its mean converges to a, i.e., $EX_n \to a$, and

(ii) its variance covariance matrix converges to zero, i.e., $VX_n \to 0$.

PROOF: This follows immediately from the equality

$$E\|X_n - a\|^2 = \operatorname{tr} VX_n + \|EX_n - a\|^2.$$

□

B.5.3 An Example of Stationary Processes: The Autoregressive Process of Order 1

Consider a univariate independent white noise $\{\epsilon_t, t \in \mathbb{Z}\}$ and a real scalar ρ smaller than one in absolute value. We are interested in the process $\{X_t, t \in \mathbb{Z}\}$, that satisfies the recursive equation

$$X_t = \rho X_{t-1} + \epsilon_t, \quad t \in \mathbb{Z} \tag{B.73}$$

Solving mechanically this equation backward gives

$$X_t = \sum_{j=0}^{\infty} \rho^j \epsilon_{t-j}. \tag{B.74}$$

It is, however, necessary to give a meaning to the above infinite sum. The generic term $\rho^j \epsilon_{t-j}$ of this series is such that

$$\sqrt{E(\rho^j \epsilon_{t-j})^2} = |\rho|^j \sqrt{E\epsilon_{t-j}^2} = |\rho|^j \sigma,$$

where σ^2 denotes the variance of ϵ_t. The series whose generic term is the above nonrandom scalar converges in the usual sense because $|\rho| < 1$. Thus, from (B.72) it follows that the infinite sum in equation (B.74) has a meaning in the sense of convergence in quadratic mean. In other words, equation (B.74) should be understood as $X_t = \lim_{J \to \infty} \sum_{j=0}^{J} \rho^j \epsilon_{t-j}$.

The fact that the preceding convergence is in quadratic mean allows us to determine easily the second-order moments of the stochastic process X_t by interverting the expectation and the summation operators. Specifically, we obtain

$$EX_t = E \sum_{j=0}^{\infty} \rho^j \epsilon_{t-j} = \sum_{j=0}^{\infty} \rho^j E\epsilon_{t-j} = 0,$$

because ϵ_t has zero mean, and

$$\begin{aligned} \text{Cov}(X_t, X_{t+h}) &= E(X_t X_{t+h}) \\ &= E\left(\sum_{j=0}^{\infty} \rho^j \epsilon_{t-j} \sum_{l=0}^{\infty} \rho^l \epsilon_{t+h-l} \right) \\ &= \sum_{j=0}^{\infty} \sum_{l=0}^{\infty} \rho^{j+l} E(\epsilon_{t-j} \epsilon_{t+h-l}). \end{aligned}$$

Now, using the independence of the ϵ_t's, we have

$$E(\epsilon_{t-j} \epsilon_{t+h-l}) = \text{Cov}(\epsilon_{t-j}, \epsilon_{t+h-l}) = 0, \text{ if } j \neq l - h.$$

Taking into account the zero terms in the expression of $\text{Cov}(X_t, X_{t+h})$, we obtain

$$\text{Cov}(X_t, X_{t+h}) = \sigma^2 \rho^h (1 + \rho^2 + \rho^4 + \ldots + \rho^{2j} + \ldots),$$

i.e.

$$\text{Cov}(X_t, X_{t+h}) = \frac{\sigma^2 \rho^h}{1 - \rho^2}. \tag{B.75}$$

Thus the stationarity property is satisfied at least for the first two moments of the process X_t. In fact, it can be proved that this process is

stationary with respect to its distribution. Moreover, if the white noise process ϵ_t is Gaussian, then the process X_t is also Gaussian since the latter is a linear transformation of the former.

B.5.4 Convergence in Distribution

The types of convergence introduced in the previous subsection apply to sequences of random vectors. On the other hand, the type of convergence studied in this subsection applies to the probability distributions of these vectors. Despite this distinction, one frequently speaks in terms of random vectors when considering convergence in distribution.

The concept of convergence in distribution has various equivalent characterizations, which correspond to the various characterizations of a probability distribution by its cumulative distribution function, its characteristic function, etc.

Definition B.32: *The sequence of random vectors X_n converges in distribution to the random vector X if either one of the following equivalent conditions is satisfied:*

(i) The cumulative distribution function F_n of X_n converges to the cumulative distribution function F of X at every continuity point of F.

(ii) The characteristic function ϕ_{X_n} of X_n converges simply to the characteristic function ϕ_X of X.

(iii) For every bounded continuous function g from $\mathbb{R}^p$ to $\mathbb{R}$, the scalar $Eg(X_n)$ converges to $Eg(X)$.

If this is the case, then we use the notation $X_n \xrightarrow{d} X$ or $X_n \approx P^X$, i.e., X_n is asymptotically distributed as P^X.

Though it deals with probability distributions, the concept of convergence in distribution is closely related to that of convergence in probability through the following properties.

Property B.59: *If X_n converges in probability to the random vector X, then X_n converges in distribution to the probability distribution of X.*

Property B.60: *If X_n converges in distribution to a nonrandom vector, i.e., if the probability distribution of X_n "converges" to the point mass distribution at a (say), then X_n converges in probability to a.*

B.5.5 Poisson Processes

We now use the preceding results on convergence in distribution so as to introduce another class of stochastic processes. We will begin with the study of the discrete time case. Then, by taking the limit, we will consider the continuous time case.

a) Discrete Case

We assume that an event A can occur at various dates indexed by $k = 1, 2, \ldots$. It is also assumed that the following conditions are satisfied:

(i) At every date, this event can either not occur or occur only once.

(ii) The fact that the event occurs at time k is independent of the fact that it has or has not occured previously.

(iii) The probability that the event occurs at date k is independent of date k. Such a probability is denoted p.

Next, we consider the date Y at which the above event occurs for the first time. We are also interested in the sequence of random variables X_k, $k = 1, 2, \ldots$, where X_k is the number of occurrences of the event of interest between dates 1 and k. It is easy to see that the random variable Y follows a Pascal distribution with parameter p, i.e.

$$P(Y = k) = (1 - p)^{k-1} p, \tag{B.76}$$

and that the random variable X_k follows a Binomial distribution $B(k, p)$.

We may equally be interested in other random variables such as the date Y_n of the n-th occurrence or the duration u_n between the $(n-1)$-th occurrence and the n-th occurrence. Note that the sequence $\{u_n, n = 1, 2, \ldots\}$ can be viewed as a random sample drawn from the Pascal distribution with parameter p. Then the random variable Y_n is obtained from $Y_n = \sum_{j=1}^{n} u_j$.

b) Continuous Case

We retain the same hypotheses as before. However, we assume that the time interval between two consecutive dates is small and equal to dt. As a matter of fact, we will consider the limit of dt to zero.

As indices of dates, we use $dt, 2dt, 3dt, \ldots, kdt, \ldots$. When the time interval dt decreases, the probability of occurrence of the event of interest

should in principle decrease. Thus we let

$$p = \lambda dt, \text{ where } \lambda \in I\!R^{+*}.$$

Now, let dt converge to zero. Depending on which variable we are considering, we will have to determine the limit of either a Pascal distribution or a Binomial distribution.

Property B.61: *When dt converges to zero, the random variable Y giving the date of the first occurrence of the event of interest is asymptotically distributed as an exponential distribution with parameter λ.*

PROOF: We establish this convergence in distribution using, for instance, the cumulative distribution function. Let Y^* denote the limit of Y_{dt}. Let $[y/dt]$ be the integer part of y/dt. We have

$$\begin{aligned}
P(Y_{dt} \leq y) &= P\left(Y_{dt} \leq \left[\frac{y}{dt}\right] dt\right) \\
&= \sum_{k=1}^{[y/dt]} (1-\lambda dt)^{k-1} \lambda dt \\
&= \lambda dt \frac{1-(1-\lambda dt)^{[y/dt]}}{\lambda dt} \\
&= 1-(1-\lambda dt)^{[y/dt]} \\
&= 1-\exp\left(\left[\frac{y}{dt}\right] \log(1-\lambda dt)\right) \\
\# \quad &1-\exp\left(\frac{y}{dt}(-\lambda dt)\right) \\
&= 1-\exp(-\lambda y),
\end{aligned}$$

which is the cumulative distribution function of the exponential distribution with parameter λ, as desired. □

Similarly, we can show that the durations u_j are independent and asymptotically distributed as an exponential distribution with parameter λ, as dt converges to zero.

We now consider the random variable X_T^* giving the number of occurrences between the dates 0 and T. The random variable X_T^* is intuitively the limit of the random variables $X_{T/dt}$, which are distributed as $B(T/dt, \lambda dt)$.

Property B.62: *When dt converges to zero, the sequence of random variables $X_{T/dt}$ converges in distribution to the random variable X_T^*, which follows the Poisson distribution with parameter λT.*

PROOF: We establish the convergence in distribution using the characteristic function. We have

$$\begin{aligned}
\phi_{dt}(u) &= E\exp(iuX_{T/dt}) \\
&= (\lambda dt \exp iu + 1 - \lambda dt)^{T/dt} \\
&\# \quad (1 - \lambda dt(1 - \exp iu))^{T/dt} \\
&\# \quad \exp\left\{\frac{T}{dt}\log(1 - \lambda dt(1 - \exp iu))\right\} \\
&\# \quad \exp -\lambda T(1 - \exp iu),
\end{aligned}$$

which is the characteristic function of the Poisson distribution $\mathcal{P}(\lambda T)$, as desired. □

It follows that the probability distribution of the number of occurrences between dates 0 and T is $\mathcal{P}(\lambda T)$.

B.6 Asymptotic Theorems

B.6.1 Laws of Large Numbers

In its simplest form the Law of Large Numbers ensures that the sample mean converges to the population mean when the observations are independently and identically distributed.

Property B.63: A Strong Law of Large Numbers. *Let $\{X_n, n \in \mathbb{N}^*\}$ be a sequence of random vectors with values in $\mathbb{R}^p$ that are independently and identically distributed with a finite vector of means. Then*

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \overset{a.s.}{\to} E(X_j),$$

as n increases to infinity.

Such a result also holds when each random variable X_j is replaced by $g(X_j)$, where g is an arbitrary function satisfying $\|Eg(X_j)\| < +\infty$. For the vectors $g(X_n)$, $n \in \mathbb{N}^*$, are also independently and identically distributed. Thus Property B.63 applies and we obtain

$$\overline{(g(X))_n} = \frac{1}{n}\sum_{j=1}^{n} g(X_j) \overset{a.s.}{\to} Eg(X_j),$$

as n increases to infinity.

In particular, we obtain the following result.

Property B.64: *Let $\{X_n, n \in \mathbb{N}^*\}$ be a sequence of random vectors with values in $\mathbb{R}^p$. Then the empirical distribution function at $(x_1, \ldots, x_p)$, i.e.*

$$F_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{X_{1j} \leq x_1, \ldots, X_{pj} \leq x_p},$$

converges almost surely to the cumulative distribution function at $(x_1, \ldots, x_p)$, i.e., to $F(x_1, \ldots, x_p) = P(X_{1j} \leq x_1, \ldots, X_{pj} \leq x_p)$.

PROOF: It suffices to define

$$g(X_j) = \mathbb{1}_{X_{1j} \leq x_1, \ldots, X_{pj} \leq x_p},$$

and to note that

$$Eg(X_j) = P(X_{1j} \leq x_1, \ldots, X_{pj} \leq x_p).$$

□

The assumptions underlying the Strong Law of Large Numbers given in Property B.63 are somewhat stringent. As a matter of fact, the random vectors X_j need be neither independent nor identically distributed. The next version, which is given for the univariate case, is useful in many situations. Its assumptions only bear on the first and second moments of X_j.

Property B.65: A Strong Law of Large Numbers. *Let $\{X_n, n \in \mathbb{N}^*\}$ be a sequence of random vectors with zero means and satisfying the following conditions:*

(i) There exists a sequence $\{r_j\}$ of real numbers between 0 and 1 such that

$$\forall\, n \geq m : \mathrm{Cov}(X_n, X_m) \leq r_{n-m} \sqrt{VX_n VX_m},$$

(ii) The sequence $\{r_j\}$ is summable, i.e., $\sum_{j=0}^{\infty} r_j < +\infty$, and

$$\sum_{j=1}^{\infty} (\log j)^2 \frac{EX_j^2}{j^2} < +\infty.$$

Then $(1/n) \sum_{j=1}^{n} X_j \xrightarrow{a.s.} 0$, as n increases to infinity.

Property B.65 is especially suited to the study of centered stationary processes. Specifically, for such processes the correlation $\mathrm{Cov}(X_n, X_m)/(\sqrt{VX_n}\sqrt{VX_m})$ depend only on the difference between n and m. Let

ρ_{n-m} denote this correlation. Then it is easy to see that the assumptions of Property B.65 hold as soon as $\sum_{j=0}^{\infty} |\rho_j| < +\infty$.

Example B.2: Consider the autoregressive of order one introduced in Section B.5.3: $X_t = \rho X_{t-1} + \epsilon_t$, where $|\rho| < 1$.

We have $EX_t = 0$ and $\text{Cov}(X_t, X_{t+h})/(\sqrt{VX_t}\sqrt{VX_{t+h}}) = \rho^{|h|}$. Because $\sum_{h=0}^{\infty} |\rho|^h < +\infty$ since $|\rho| < 1$, it follows that $(1/T)\sum_{t=1}^{T} X_t \xrightarrow{a.s.} 0$, as T increases to infinity.

B.6.2 Central Limit Theorems

Let $\{X_n, n \in I\!N^*\}$ be a sequence of random vectors independently and identically distributed with mean vector m and variance covariance matrix Σ. Then the Law of Large Numbers given in Property B.63 ensures that $\bar{X}_n - m$ converges almost surely to zero. However, we may want to know the rate of convergence to zero of this difference. In particular, we may want to find a quantity that is asymptotically equivalent to this difference.

Theorem B.8: Central Limit Theorem. *Under the above assumptions, the sequence* $\sqrt{n}(\bar{X}_n - m)$ *converges in distribution to the normal distribution* $N(0, \Sigma)$, *i.e.,* $\sqrt{n}(\bar{X}_n - m) \xrightarrow{d} N(0, \Sigma)$.

PROOF: This result follows at once from the study of the characteristic function of $Y_n = \sqrt{n}(\bar{X}_n - m)$. Namely, we have

$$\begin{aligned}
\phi_{Y_n}(u) &= E(\exp iu'Y_n) \\
&= E\left(\exp i\frac{u'}{\sqrt{n}}\sum_{j=1}^{n}(X_j - m)\right) \\
&= \prod_{j=1}^{n} E\left(\exp i\frac{u'}{\sqrt{n}}(X_j - m)\right) \\
&= \left(E\left(\exp i\frac{u'}{\sqrt{n}}(X_j - m)\right)\right)^n,
\end{aligned}$$

where the third and fourth equalities follow from the independence and the identical distribution assumptions, respectively. Hence, when n increases to infinity, we have

$$\left(E\left(\exp i\frac{u'}{\sqrt{n}}(X_j - m)\right)\right)^n$$

$$\begin{aligned} & \left(E\left(1+\frac{iu'}{\sqrt{n}}(X_j-m)-\frac{1}{2n}(u'(X_j-m))^2\right)\right)^n \\ = & \left(1-\frac{1}{2n}u'\Sigma u\right)^n \\ \# & \exp -\frac{1}{2}u'\Sigma u, \end{aligned}$$

which is the characteristic function of the $N(0,\Sigma)$ distribution. □

Example B.3: Let $\{X_n, n \in I\!N^*\}$ be a sequence of random variables that are independently and identically distributed as a Bernoulli distribution $B(1,p)$. We know that $\sum_{j=1}^n X_j$ follows a Binomial distribution $B(n,p)$.

Now, applying the preceding Central Limit Theorem shows that the probability distribution of $\sqrt{n}(\bar{X}_n - p)$ approaches the normal distribution $N(0, p(1-p))$, as n increases to infinity. Equivalently, using the fact that the family of normal distributions is closed under linear transformations, we obtain that the probability distribution of $\sum_{j=1}^n X_j$ can be approximated by the normal distribution $N(np, np(1-p))$. In other words, the Binomial distribution $B(n,p)$ can be approximated by the normal distribution $N(np, np(1-p))$, when n is large.

Example B.4: Assume that the assumptions of Property B.64 hold. Then the mean of the random variables

$$g(X_j) = I\!I_{X_{1j} \leq x_1, \ldots, X_{pj} \leq x_p}$$

is $F(x_1, \ldots, x_p)$, while their variance is $F(x_1, \ldots, x_p)(1 - F(x_1, \ldots, x_p))$. From the preceding Central Limit Theorem it follows that

$$\begin{aligned} & \sqrt{n}(F_n(x_1, \ldots, x_p) - F(x_1, \ldots, x_p)) \\ & \quad \xrightarrow{d} N(0, F(x_1, \ldots, x_p)(1 - F(x_1, \ldots, x_p))). \end{aligned}$$

As for the Law of Large Number given in Property B.63, the assumptions underlying the Central Limit Theorem given in Theorem B.8 are quite stringent. However, these assumptions can be greatly weakened. For instance, we have the following result:

Property B.66: *Let* $\{X_t, t \in \mathbf{Z}\}$ *be a univariate stationary process defined by*

$$X_t = \sum_{j=0}^{\infty} a_j \epsilon_{t-j}, \text{ with } \sum_{j=0}^{\infty} |a_j| < \infty, E\epsilon_t^4 < \infty,$$

and $\{\epsilon_t, t \in \mathbf{Z}\}$ *is a white noise process. Then* $\sqrt{T}\bar{X}_T = (\sqrt{T})^{-1}\sum_{t=1}^{T} X_t$ *converges in distribution to the centered normal distribution with variance* $\sum_{h=-\infty}^{+\infty} \mathrm{Cov}(X_t, X_{t+h})$, *i.e.*

$$\sqrt{T}\bar{X}_T \approx N\left(0, \sum_{h=-\infty}^{+\infty} \mathrm{Cov}(X_t, X_{t+h})\right).$$

Note that asymptotic normality still holds even though independence was not required. Note, however, that the asymptotic variance now takes into account the existing correlations between the variables X_j.

Example B.5: Consider again the autoregressive process of order one. We have

$$\mathrm{Cov}(X_t, X_{t+h}) = \frac{\sigma^2}{1-\rho^2}\rho^{|h|}.$$

Hence

$$\begin{aligned}\sum_{h=-\infty}^{+\infty} \mathrm{Cov}(X_t, X_{t+h}) &= \frac{\sigma^2}{1-\rho^2}\left(2\sum_{h=0}^{\infty}\rho^h - 1\right) \\ &= \frac{\sigma^2}{1-\rho^2}\left(\frac{2}{1-\rho} - 1\right) \\ &= \frac{\sigma^2}{(1-\rho)^2}.\end{aligned}$$

Thus

$$\sqrt{T}\bar{X}_T \xrightarrow{d} N\left(0, \frac{\sigma^2}{(1-\rho)^2}\right).$$

It is frequently useful to know the (joint) asymptotic distribution of the sample autocorrelations of a stationary process

$$\hat{\rho}(h) = \frac{\sum_{t=1}^{T-|h|}(X_t - \bar{X}_T)(X_{t+|h|} - \bar{X}_T)}{\sum_{t=1}^{T}(X_t - \bar{X}_T)^2}.$$

Property B.67: *Under the assumptions of Property B.66 (including* $E\epsilon_t^4 < \infty$*), we have*

$$\sqrt{T}(\hat{\rho}(1) - \rho(1), \ldots, \hat{\rho}(k) - \rho(k)) \xrightarrow{d} N(0, \mathbf{Q}),$$

for every k, where $\mathbf{Q}$ is the matrix whose (h,l) element is

$$\sum_{j=-\infty}^{+\infty} (\rho(j)\rho(j-h+l) + \rho(j+l)\rho(j-h) - 2\rho(l)\rho(j)\rho(j-h) \\ -2\rho(h)\rho(j)\rho(j-l) + 2\rho(h)\rho(l)\rho^2(j)).$$

Because $\bar{X}_T$ converges almost surely to zero, we note that the preceding property still holds when $\hat{\rho}(h)$ is replaced by

$$\tilde{\rho}(h) = \frac{\sum_{t=1}^{T-|h|} X_t X_{t+|h|}}{\sum_{t=1}^{T} X_t^2}.$$

Example B.6: Consider again the autoregressive process of order one. The sample autocorrelation of order h is

$$\tilde{\rho}(h) = \frac{\sum_{t=1}^{T-|h|} X_t X_{t+|h|}}{\sum_{t=1}^{T} X_t^2}.$$

Then Property B.67 implies that $\sqrt{T}(\tilde{\rho}(h) - \rho^{|h|})$ converges in distribution to a centered normal distribution with a variance given by the above formula with $l = h$ and $\rho(j) = \rho^{|j|}$. Specifically, we obtain

$$\sum_{j=-\infty}^{+\infty} \left[\rho^{2|j|} + \rho^{|j+h|+|j-h|} - 4\rho^{|h|}\rho^{|j|+|j-h|} + 2\rho^{2|h|}\rho^{2|j|}\right],$$

i.e.

$$\frac{(1+\rho^2)(1-\rho^{2h})}{1-\rho^2} - 2h\rho^{2h},$$

where $h > 0$. In particular the asymptotic variance of $\sqrt{T}(\tilde{\rho}(1) - \rho)$ is

$$1 + \rho^2 - 2\rho^2 = 1 - \rho^2.$$

B.6.3 Applications to Continuous Differentiable Functions

The preceding results apply not only to the asymptotic behavior of sample means but also to that of continuously differentiable functions of

these sample means. Specifically, it suffices to invoke the following properties.

Property B.68: *Let $\{Y_n, n \in \mathbb{N}^*\}$ be a sequence of random vectors with values in $\mathbb{R}^p$ converging almost surely (or in probability) to the random vector X. Let g be a continuous function from $\mathbb{R}^p$ to $\mathbb{R}^q$. Then the sequence $\{g(X_n)\}$ converges almost surely (or in probability) to $g(X)$.*

Example B.7: Let $\{X_n, n \in \mathbb{N}^*\}$ be a sequence of independently and identically distributed random variables with finite fourth moments. Then, from the Law of Large Numbers we know that

$$\begin{pmatrix} \bar{X}_n \\ \overline{(X^2)}_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{j=1}^n X_j \\ \sum_{j=1}^n X_j^2 \end{pmatrix}.$$

converges almost surely to $(E(X_j)\ E(X_j^2))'$. Hence $\overline{(X^2)}_n - (\bar{X}_n)^2$ converges almost surely to $E(X_j^2) - (EX_j)^2$, i.e., the sample variance converges almost surely to the population variance.

Property B.69: *Let $\{Y_n, n \in \mathbb{N}^*\}$ be a sequence of random vectors in $\mathbb{R}^p$ that are asymptotically normally distributed in the sense that $\sqrt{n}(Y_n - m) \xrightarrow{d} N(0, \Sigma)$. If g is a continuously differentiable function from $\mathbb{R}^p$ to $\mathbb{R}^q$, then*

$$\sqrt{n}(g(Y_n) - g(m)) \xrightarrow{d} N\left(0, \frac{\partial g(m)}{\partial y'} \Sigma \frac{\partial g'(m)}{\partial y}\right),$$

where $\partial g(y)/\partial y'$ denotes the $q \times p$ Jacobian matrix of g.

The proof is quite simple if we admit thtat $g(Y_n)$ can be expanded in a Taylor series around m. Then in this case we have

$$\sqrt{n}(g(Y_n) - g(m)) \# \frac{\partial g(m)}{\partial y'} \sqrt{n}(Y_n - m).$$

Thus the left-hand side is asymptotically equivalent to a linear transformation of a random vector that is asymptotically normally distributed. The asymptotic normality of $\sqrt{n}(g(Y_n) - g(m))$ as well as its asymptotic variance covariance matrix follow immediately.

B.6.4 Other Asymptotic Theorems

It is possible to obtain more precise results on the asymptotic behavior of the sample mean. We give below some classical results, where we restrict ourselves to the case of scalar random variables X_n, $n \in \mathbb{N}^*$, that are independently and identically distributed, to simplify.

a) Glivenko–Cantelli Theorem

We have mentioned that the empirical cumulative distribution function $F_n(x)$ at x converges almost surely to the true cumulative distribution function at x, for every x. Using, however, the monotony of the cumulative distribution function, it can be proved that such a convergence is uniform.

Theorem B.9: Glivenko–Cantelli Theorem. *Under the preceding assumptions,* $\sup_{-\infty<x<+\infty} |F_n(x) - F(x)|$ *converges almost surely to zero as* n *increases to infinity.*

b) The Law of the Iterated Logarithm

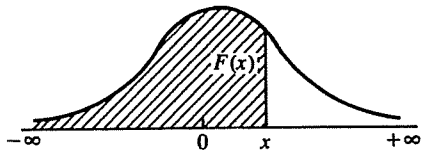
From the Central Limit Theorem, it follows in particular that $\sqrt{n}(\bar{X}_n - m)$ is between -1.96σ and 1.96σ with probability approaching 0.95. This result, however, does not say much about the extreme values of $\sqrt{n}(\bar{X}_n - m)$. This is the purpose of the next property.

Property B.70: Law of the Iterated Logarithm. *We have almost surely*

$$\limsup_n \frac{\sqrt{n}(\bar{X}_n - m)}{\sigma\sqrt{2\log\log n}} = 1 \quad \textit{and} \quad \liminf_n \frac{\sqrt{n}(\bar{X}_n - m)}{\sigma\sqrt{2\log\log n}} = -1.$$

Table 1

Cumulative Distribution Function of the Standard Normal Distribution

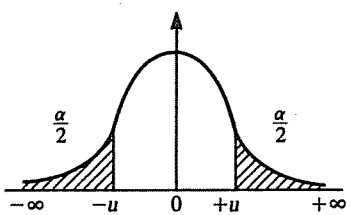


$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500 00	0.503 99	0.507 98	0.511 97	0.515 95	0.519 94	0.523 92	0.527 90	0.531 88	0.535 86
0.1	0.539 83	0.543 80	0.547 76	0.551 72	0.555 67	0.559 62	0.563 56	0.567 50	0.571 42	0.575 35
0.2	0.579 26	0.583 17	0.587 06	0.590 95	0.594 84	0.598 71	0.602 57	0.606 42	0.610 26	0.614 09
0.3	0.617 91	0.621 72	0.625 52	0.629 30	0.633 07	0.636 83	0.640 58	0.644 31	0.648 03	0.651 73
0.4	0.655 42	0.659 10	0.662 76	0.666 40	0.670 03	0.673 65	0.677 24	0.680 82	0.684 39	0.687 93
0.5	0.691 46	0.694 97	0.698 47	0.701 94	0.705 40	0.708 84	0.712 26	0.715 66	0.719 04	0.722 40
0.6	0.725 75	0.729 07	0.732 37	0.735 65	0.738 91	0.742 15	0.745 37	0.748 57	0.751 75	0.754 90
0.7	0.758 04	0.761 15	0.764 24	0.767 31	0.770 35	0.773 37	0.776 37	0.779 35	0.782 30	0.785 24
0.8	0.788 14	0.791 03	0.793 89	0.796 73	0.799 55	0.802 11	0.805 11	0.807 85	0.810 57	0.813 27
0.9	0.815 94	0.818 59	0.821 21	0.823 81	0.826 39	0.828 94	0.831 47	0.833 98	0.836 46	0.838 91
1.0	0.841 34	0.843 75	0.846 14	0.848 50	0.850 83	0.853 14	0.855 43	0.857 69	0.859 93	0.862 14
1.1	0.864 33	0.866 50	0.868 64	0.870 76	0.872 86	0.874 93	0.876 98	0.879 00	0.881 00	0.882 98
1.2	0.884 93	0.886 86	0.888 77	0.890 65	0.892 51	0.892 35	0.896 17	0.897 96	0.899 73	0.901 47
1.3	0.903 20	0.904 90	0.906 58	0.908 24	0.909 88	0.911 49	0.913 09	0.914 66	0.916 21	0.917 74
1.4	0.919 24	0.920 73	0.922 20	0.923 64	0.925 07	0.926 47	0.927 86	0.929 22	0.930 56	0.931 89
1.5	0.933 19	0.934 48	0.935 74	0.936 99	0.938 22	0.939 43	0.940 62	0.941 79	0.942 95	0.944 08
1.6	0.945 20	0.946 30	0.947 38	0.948 45	0.949 50	0.950 53	0.951 54	0.952 54	0.953 52	0.954 49
1.7	0.955 43	0.956 37	0.957 28	0.958 19	0.959 07	0.959 94	0.960 80	0.961 64	0.962 46	0.963 27
1.8	0.964 07	0.964 85	0.965 62	0.966 38	0.967 12	0.967 84	0.968 56	0.969 26	0.969 95	0.970 62
1.9	0.971 28	0.971 93	0.972 57	0.973 20	0.973 81	0.974 41	0.975 00	0.975 58	0.976 15	0.976 70
2.0	0.977 25	0.977 78	0.978 31	0.978 82	0.979 32	0.979 82	0.980 30	0.980 77	0.981 24	0.981 69
2.1	0.982 14	0.982 57	0.983 00	0.983 41	0.983 82	0.984 22	0.984 61	0.985 00	0.985 37	0.985 74
2.2	0.986 10	0.986 45	0.986 79	0.987 13	0.987 45	0.987 78	0.988 09	0.988 40	0.988.70	0.988 99
2.3	0.989 28	0.989 56	0.989 83	0.990 10	0.990 36	0.990 86	0.991 11	0.991 11	0.991 34	0.991 58
2.4	0.991 80	0.992 02	0.992 24	0.992 45	0.992 66	0.992 86	0.993 05	0.993 24	0.993 43	0.993 61
2.5	0.993 79	0.993 96	0.994 13	0.994 30	0.994 46	0.994 61	0.994 77	0.994 92	0.995 06	0.995 20
2.6	0.995 34	0.995 47	0.995 60	0.995 73	0.995 85	0.995 98	0.996 09	0.996 21	0.996 32	0.996 43
2.7	0.996 53	0.996 64	0.996 74	0.996 83	0.996 93	0.997 02	0.997 11	0.997 20	0.997 28	0.997 36
2.8	0.997 44	0.997 52	0.997 60	0.997 67	0.997 74	0.997 81	0.997 88	0.997 95	0.998 01	0.998 07
2.9	0.998 13	0.998 19	0.998 25	0.998 31	0.998 36	0.998 41	0.998 46	0.998 51	0.998 56	0.998 61

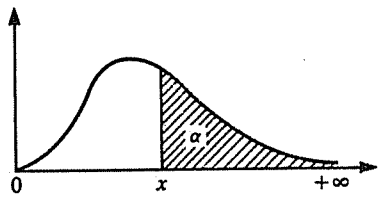
Table 2

Quantiles of the Standard Normal Distribution (u = value of Z such that $Pr(\mid Z \mid > u) = \alpha$)



α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	2.5758	2.3263	2.1701	2.0537	1.9600	1.8808	1.8119	1.7507	1.6954
0.1	1.6449	1.5982	1.5548	1.5141	1.4758	1.4395	1.4051	1.3722	1.3408	1.3106
0.2	1.2816	1.2536	1.2566	1.2004	1.1750	1.1503	1.1264	1.1031	1.0803	1.0581
0.3	1.0364	1.0152	0.9945	0.9741	0.9542	0.9346	0.9154	0.8965	0.8779	0.8596
0.4	0.8416	0.8239	0.7892	0.8064	0.7722	0.7554	0.7388	0.7255	0.7063	0.6903
0.5	0.6745	0.6588	0.6433	0.6280	0.6128	0.5978	0.5828	0.5681	0.5534	0.5388
0.6	0.5244	0.5101	0.4959	0.4817	0.4677	0.4538	0.4399	0.4261	0.4125	0.3989
0.7	0.3853	0.3719	0.3585	0.3451	0.3319	0.3186	0.3000	0.2924	0.2793	0.2663
0.8	0.2533	0.2404	0.2275	0.2147	0.2019	0.1819	0.1764	0.1637	0.1510	0.1383
0.9	0.1257	0.1130	0.1004	0.0878	0.0753	0.0627	0.0502	0.0376	0.0251	0.0125

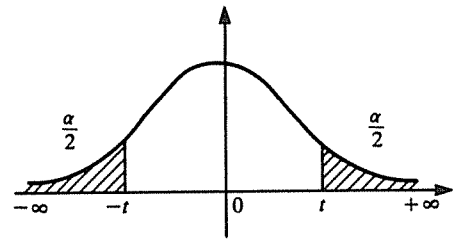
Table 3
Quantiles of the Chi–Square Distribution with ν Degrees of Freedom
(x = value of χ^2 such that $Pr(\chi^2 > x) = \alpha$)



ν \ α	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.001
1	0.000 2	0.001 0	0.003 9	0.015 8	2.71	3.84	5.02	6.63	10.83
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	13.82
3	0.12	0.22	0.35	0.58	6.25	7.81	9.35	11.34	16.27
4	0.30	0.48	0.71	1.06	7.78	9.94	11.14	13.28	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	20.52
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.47	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	26.13
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	29.59
11	3.05	3.82	4.57	5.58	17.27	19.67	21.92	24.72	31.26
12	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	32.91
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	34.53
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	36.12
15	5 .23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	37.70
16	5.81	6.91	7.96	9.31	23.54	26.30	28.84	32.00	39.25
17	6.41	7.56	8.67	10.08	24.77	27.59	30.19	33.41	40.79
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.80	42.31
19	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	43.82
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	45.32
21	8.90	10.28	11.59	13.24	29.61	32.67	35.48	38.93	46.80
22	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	48.27
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	49.73
24	10.86	12.40	13.85	15.66	33.20	36.41	39.37	42.98	51.18
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
26	12.20	13.84	15.38	17.29	35.56	38.88	41.92	45.64	54.05
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	55.48
28	13.57	15.31	16.93	18.94	37.92	41.34	44.46	48.28	56.89
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	58.30
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	59.70

when $\nu > 30$. then $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ is approximately $N(0.1)$.

Table 4
Quantiles of the Student Distribution with ν Degrees of Freedom
(t = value of T such that $Pr(|\ T\ |> t) = \alpha$)

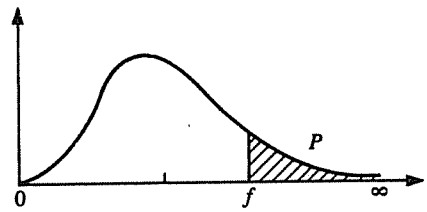


ν \ α	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.929
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.533	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.377	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.137	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.649
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.656
40	0.127	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
80	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Table 5

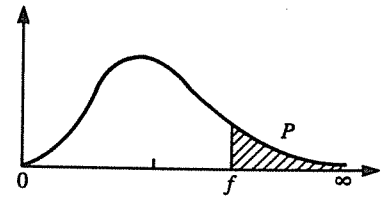
Quantiles of the Fisher–Snedecor Distribution with ν_1 and ν_2 Degrees of Freedom

(f = value of F such that $Pr(F > f) = \alpha$)



	$\nu_1 = 1$		$\nu_1 = 2$		$\nu_1 = 3$		$\nu_1 = 4$		$\nu_1 = 5$	
ν_2	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$
1	161.4	4 052	199.5	4 999	215.7	5 403	224.6	5 625	230.2	5 764
2	18.51	98.49	19.00	99.00	19.16	99.17	19.25	99.25	19.30	99.30
3	10.13	34.12	9.55	30.81	9.28	29.46	9.12	28.71	9.01	28.24
4	7.71	21.20	6.94	18.00	6.59	16.69	6.39	15.98	6.26	15.52
5	6.61	16.26	5.79	13.27	5.41	12.06	5.19	11.39	5.05	10.97
6	5.99	13.74	5.14	10.91	4.76	9.78	4.53	9.15	4.39	8.75
7	5.59	12.15	4.74	9.55	4.35	8.45	4.12	7.85	3.97	7.45
8	5.32	11.26	4.46	8.65	4.07	7.59	3.84	7.01	3.69	6.63
9	5.12	10.56	4.26	8.02	3.86	6.99	3.63	6.42	3.48	6.06
10	4.96	10.04	4.10	7.56	3.71	6.55	3.48	5.99	3.33	5.64
11	4.84	9.65	3.98	7.20	3.59	6.22	3.36	5.67	3.20	5.32
12	4.75	9.33	3.88	6.93	3.49	5.95	3.26	5.41	3.11	5.06
13	4.67	9.07	3.80	6.70	3.41	5.74	3.18	5.20	3.02	4.86
14	4.60	8.86	3.74	6.51	3.34	5.56	3.11	5.03	2.96	4.69
15	4.54	8.86	3.68	6.36	3.29	5.42	3.06	4.89	2.90	4.56
16	4.49	8.53	3.63	6.23	3.24	5.29	3.01	4.77	2.85	4.44
17	4.45	8.40	3.59	6.11	3.20	5.18	2.96	4.67	2.81	4.43
18	4.41	8.28	3.55	6.01	3.16	5.09	2.93	4.58	2.77	4.25
19	4.38	8.18	3.52	5.93	3.13	5.01	2.90	4.50	2.74	4.17
20	4.35	8.10	3.49	5.85	3.10	4.94	2.87	4.43	2.71	4.10
21	4.32	8.02	3.47	5.78	3.07	4.87	2.84	4.37	2.68	4.04
22	4.30	7.94	3.44	5.72	3.05	4.82	2.82	4.31	2.66	3.99
23	4.28	7.88	3.42	5.66	3.03	4.76	2.80	4.26	2.64	3.94
24	4.26	7.82	3.40	5.61	3.01	4.72	2.78	4.22	2.62	3.90
25	4.24	7.77	3.38	5.57	2.09	4.68	2.76	4.18	2.60	3.86
26	4.22	7.72	3.37	5.53	2.98	4.64	2.74	4.14	2.59	3.82
27	4.21	7.68	3.35	5.49	2.96	4.60	2.73	4.11	2.57	3.70
28	4.20	7.64	3.34	5.45	2.95	4.57	2.71	4.07	2.56	3.75
29	4.18	7.60	3.33	5.42	2.93	4.54	2.70	4.04	2.54	3.73
30	4.17	7.56	3.32	5.39	2.92	4.51	2.69	4.02	2.53	3.70
40	4.08	7.31	3.23	5.18	2.84	4.31	2.61	3.83	2.45	3.51
60	4.00	7.08	3.15	4.98	2.76	4.13	2.52	3.65	2.37	3.34
120	3.92	6.85	3.07	4.79	2.68	3.95	2.45	3.48	2.29	3.17
∞	3.84	6.64	2.99	4.60	2.60	3.78	2.37	3.32	2.21	3.02

Table 5 Continued



	$\nu_1 = 6$		$\nu_1 = 8$		$\nu_1 = 12$		$\nu_1 = 24$		$\nu_1 = \infty$	
ν_2	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$	$P = 0.05$	$P = 0.01$
1	234.0	5 859	238.9	5 981	243.9	6 106	249.0	6 234	254.3	6 366
2	19.33	99.33	19.37	99.36	19.41	99.42	19.45	99.46	19.50	99.50
3	8.94	27.91	8.84	27.49	8.74	27.05	8.64	26.60	8.53	26.12
4	6.61	15.21	6.04	14.80	5.91	14.37	5.77	13.93	5.63	13.46
5	4.95	10.67	4.82	10.27	4.68	9.89	4.53	9.47	4.36	9.02
6	4.28	8.47	4.15	8.10	4.00	7.72	3.84	7.31	3.67	6.88
7	3.87	7.19	3.73	6.84	3.57	6.47	3.41	6.07	3.23	5.65
8	3.58	6.37	3.44	6.03	3.28	5.67	3.12	5.28	2.93	4.86
9	3.37	5.80	3.23	5.47	3.07	5.11	2.90	4.73	2.71	4.31
10	3.22	5.39	3.07	5.06	2.91	4.71	2.74	4.33	2.54	3.91
11	3.09	5.07	2.95	4.74	2.79	4.40	2.61	4.02	2.40	3.60
12	3.00	4.82	2.85	4.50	2.69	4.16	2.50	3.78	2.30	3.36
13	2.92	4.62	2.77	4.30	2.60	3.96	2.42	3.59	2.21	3.16
14	2.85	4.46	2.70	4.14	2.53	3.80	2.35	3.43	2.13	3.00
15	2.79	4.32	2.64	4.00	2.48	3.67	2.29	3.29	2.07	2.87
16	2.74	4.20	2.59	3.89	2.42	3.55	2.24	3.18	2.01	2.75
17	2.70	4.10	2.55	3.79	2.35	3.45	2.19	3.08	1.96	2.65
18	2.66	4.01	2.51	3.71	2.34	3.37	2.15	3.00	1.92	2.57
19	2.63	3.94	2.48	3.63	2.31	3.30	2.11	2.92	1.88	2.49
20	2.60	3.87	2.45	3.56	2.28	3.23	2.68	2.86	1.84	2.42
21	2.57	3.81	2.42	3.51	2.25	3.17	2.05	2.80	1.81	2.36
22	2.55	3.76	2.40	3.45	2.23	3.12	2.03	2.75	1.78	2.31
23	2.53	3.71	2.38	3.41	2.20	3.07	2.00	2.70	1.76	2.26
24	2.51	3.67	2.36	3.36	2.18	1.03	1.98	2.66	1.73	2.21
25	2.49	3.63	2.34	3.32	2.16	2.99	1.96	2.62	1.71	2.17
26	2.47	3.59	2.32	3.29	2.15	2.96	1.95	2.58	1.69	2.13
27	2.46	3.56	2.30	3.26	2.13	2.93	1.93	2.55	1.67	2.10
28	2.44	3.53	2.29	3.23	2.12	2.90	1.91	2.52	1.65	2.06
29	2.43	3.50	2.28	3.20	2.10	2.87	1.90	2.49	1.64	2.03
30	2.42	3.47	2.27	3.17	2.09	2.84	1.89	2.47	1.62	2.01
40	2.34	3.29	2.18	2.99	2.00	2.66	1.79	2.29	1.51	1.80
60	2.25	3.12	2.10	2.82	1.92	2.50	1.70	2.12	1.39	1.60
120	2.17	2.96	2.01	2.66	1.83	2.34	1.61	1.95	1.25	1.38
∞	2.09	2.80	1.94	2.51	1.75	2.18	1.52	1.79	1.09	1.00

Index

Acceptance region, 221
Admissible Tests, 11, 21
Akaike information criterion, 307
α-similar tests
  Confidence region and, 222
  Definition, 42
  Neyman α-structure and, 48
  Unbiased tests and, 42
  Uniformly most powerful tests and, 42
Asymptotically efficient semiparametric estimators, 376, 379
Asymptotic bias, 337
Asymptotic comparison of estimators, 325
Asymptotic comparison of tests
  Approximate methods, 346, 352, 355
  Direct methods, 347, 351, 355, 366
  Dual methods, 349, 351
  Exact methods, 346, 351
Asymptotic efficiency of estimators
  Definition, 326, 327, 331
  First-order, 326
  Newton–Raphson algorithm and, 342, 379
  Second-order, 341
  Semiparametric, 370
  Superefficiency, 334
Asymptotic least squares estimation, see also Extremum estimation
  General asymptotic tests and, 148
  Inequality constraints and, 243
  Specification tests and, 153
Asymptotic relative efficiency of tests
  Bahadur, 354
  Pitman, 362
Asymptotic tests, see also Likelihood based tests, General asymptotic tests
  Consistency, 84
  Definition, 84
Atkinson comprehensive model, 298
Augmented regression
  Exogeneity tests and, 167
  Score tests and, 114, 290
Autocorrelation tests, 117, 129, 132, 207, 209
Autoregressive process
  First-order, 499, 506, 508, 509
  Order k (of), 117, 159

Bahadur approach, 351, 354
Bases
  Change of, 411
  Definition, 406
  Orthonormal, 423, 433
Bayesian
  Confidence regions, 228
  Model selection criteria, 312
  Principle, 13
  Risk, 13, 21, 312
  Tests, 22
Bienaymé–Tchebychev inequality, 499
Bijective, 406
Bilinear form, 420
Bonferroni Inequality, 203
Box–Cox comprehensive model, 300
Box–Cox transformation, 300
Break, 62

$C(\alpha)$ tests
  Asymptotic equivalent tests, 123
  Definition, 123
  R^2 and, 124
Canonical parameters, 48, 54
Cauchy–Schwarz inequality, 469
Causality measures, 280, 320
Central limit theorems, 363, 506
Characteristic equation, 412
Characteristic function, 449, 457, 472
Chi-square tests, see Goodness-of-fit tests, Independence tests
Chow test, 62
Cobb–Douglas Production function, 3
Cochran theorem, 480, 490

Coefficient of determination
  Adjusted, 305, 322
  Definition, 304
  Inconsistency of $\bar{R}^2$, 306
  Score tests and, 90, 100, 103, 115, 116, 117, 118
Common root tests, 200
Comprehensive model, 297
Conditional median model, 375, 377
Confidence intervals, 214, 215, 223, 224
Confidence level, 213
Confidence regions
  Asymptotic, 224, 227
  Bayesian, 228
  Conservative, 218
  Definition, 213
  Linear models and, 217, 219
  Optimality (of), 219
  Tests and, 222
  Unbiased, 220, 223
  Uniformly most precise, 220
  Uniformly most precise and unbiased, 221, 223
Contingency tables, 63
Convergence
  Almost sure, 386, 498, 510
  In distribution, 501, 510
  In probability, 398, 498, 501, 510
  In quadratic mean, 498, 499
  Uniform almost sure, 386, 389
Correlation coefficients
  Empirical, 118, 131
  Empirical partial, 183
  Multiple, 305
Correlation matrix, 456
Covariance matrix
  Conditional, 464
  Conditional variance, 463, 482
  Definition, 455
  Partial, 468
  Partial variance, 468
  Variance, 455, 463
Cox tests, 293, 317
Critical regions
  Approximate, 352
  Confidence regions and, 221
  Definition, 7
  Nonrandomized tests and, 8
  Score tests and, 33
Cumulative distribution function, 445

Davidson–McKinnon comprehensive model, 301
Decision rules
  Definition, 6
  Nonrandomized, 7
  Randomized, 7
Density functions, 447
Disequilibrium models, 235, 273
Dual space, 406
Durbin–Watson statistic, 119

Eigenvalue, 412
Eigenvector, 412
Error independence test, 117
Error type, 7
Estimating equations, 160
Euclidean norm, 421
Exogeneity tests, 167
Expansions
  Edgeworth, 343
  High-order, 334
  Taylor series, 337, 397
Expectation
  Conditional, 461, 462
  Linear regression and conditional, 465, 467, 481
  Random variables (of), 449, 450, 455
  Random vectors (of), 451
  Rational, 5
Exponential models
  Latent, 119
  Neyman α-structure and, 48
  Neyman tests and, 26
  One-sided hypothesis tests and, 30, 49
  Simple hypothesis tests and, 26
  Two-sided hypothesis and, 36, 52
  Uniformly most powerful tests in, 26, 30, 36, 43, 49, 52, 54, 77
Extremum estimation, see also Inequality constraints
  Asymptotic normality, 401
  Consistency, 386
  Constrained, 395
  Existence, 383
  Quasi generalized, 392, 394
  Unconstrianed, 394

Fisher information
  Conditional, 83, 164
  Definition, 33
  Estimation, 34, 83, 85, 89
  Nonlinear models and, 101
Fisher statistic, 75, 99, 181, 305
Fisher tests, 72, 180, 305
Fisher theorem, 481, 491
Frisch–Waugh theorem, 435, 468

General asymptotic tests, see also Inequality constraints
Asymptotic equivalence, 141, 142, 146, 147, 148, 150, 250, 251, 253
Asymptotic least squares estimation (using), 148
Based on $g(\hat{\theta}_n, a)$, 147
Constrained maximization (using), 136
Hausman–Wald tests, 139, 144, 146, 147, 249, 260
Invariance, 145
Lagrange multiplier tests, 139, 144, 146, 147, 152
Least squares estimation and, 150
Likelihood ratio type tests, 142, 248, 260, 266
Score tests, 140, 144, 146, 147, 252, 260
Unconstrained estimator (using), 145
Wald type tests, 138, 144, 146, 147, 251, 260, 262, 266
When $\mathcal{I}_0 = \mathcal{J}_0$, 143, 146
Generalized methods of moments, see also Extremum estimation
Definition, 160
Inequality constraints and, 244
Specification tests, and, 160
Generalized residuals
Definition, 120
Score tests and, 119
Glivenko–Cantelli theorem, 511
Goodness-of-fit tests
Distribution (to a), 113
Family of distributions (to a), 109

Hausman test, see also General asymptotic tests
Asymptotic equivalent, 126
Definition, 125
Specification tests and, 166
Statistic, 168
Hermite polynomial, 344
Homoscedasticity tests, 116, 131, 208, 209, 211
Hypothesis
Alternative, 2
Composite, 2
Explicit form, 103, 109, 198
General, 2, 184, 198
Implicit form, 82, 96, 184, 195
Invariant, 70
Linear, 72
Maintained, see General
Mixed form, 136, 174
Nested, 179, 195, 200, 275
Nonidentified, 5, 6
Nonnested, 202, 282
Null, 2, 15, 82, 96, 103, 109, 184, 195, 198
One-sided, 28, 30, 49
Scalar linear, 54, 62
Simple, 2, 17
Two-sided, 35, 43, 51

Idempotent, 431
Identified
Asymptotically, 89, 107
First-order, 5
Model, 2, 70
Testing problem, 3, 5, 6
Image of a linear mapping, 406, 423
Independence tests, 63, 111
Individual effects
Definition, 263
Test for, 264
Inequality constraints
Asymptotically equivalent tests, 250, 251, 253, 255
Asymptotic properties of estimators under, 242, 246
Dual tests, 255, 260, 264
Estimation under, 237, 244
Hausman–Wald type tests, 250, 260, 266
Kühn and Tücker multipliers tests, 250, 260
Likelihood ratio type tests and, 248, 260, 266
Linear models and, 237, 245, 261, 271, 272
Score tests, 252, 260
Tests of, 266
Tests under maintained, 247
Wald type tests, 251, 260, 262
Infinitely small quantities, 398
Information matrix test, 170
Injective, 406
Instrumental variables validity test, 157, 163
Invariance
Asymptotic, 145
Exact, 145
Principle, 14, 70
Invariant tests
Definition, 71
Maximal, 71

Uniformly most powerful and, 71, 75
Inverse
Generalized, 437
Mapping, 406, 438
Matrix, 410
Moore–Penrose, 438
Partitioned matrix, 417

Jacknife principle, 340
Jensen inequality, 449, 463
J test, 303

Kernel of a linear mapping, 406, 423
Koyck type hypothesis, 156
Kühn and Tücker multipliers, 239, 245
Kühn and Tücker multiplier tests, 250, 260
Kullback information, 276
Kullback proximity, 276, 278, 281

Lagrange multipliers, 87, 137, 185, 240, 254, 395
Lagrange multiplier tests, see Score tests
Laplace transform, 457
Law of the iterated logarithm, 511
Laws of large number, 504
Least squares estimation, see also Asymptotic least squares estimation
Generalized, 152
Inequality constraints and, 237, 245
Ordinary, 58, 151, 168, 339
Quasi generalized, 377
Two-stage, 158, 163, 168
Likelihood Principle, 14
Likelihood based tests, see $C(\alpha)$-tests, Hausman test, Likelihood ratio tests, Score tests, Wald tests
Likelihood ratio, 23, 25, 28
Likelihood ratio tests, see also General asymptotic tests
Asymptotically equivalent tests, 91, 93, 107, 108, 123, 126, 185, 191, 193
$C(\alpha)$ tests and, 123
Consistency, 91, 107
Definition, 91
Finite sample distribution, 99
Hausman tests and, 126
Independence property, 94, 98, 109, 194
Nonnested hypotheses and, 293
Null hypothesis in explicit form and, 107
Null hypothesis in implicit form and, 90, 184
Principle, 81
Regression models and, 97, 100
Score tests and, 91, 100, 107, 185
Wald tests and, 91, 100, 107, 185
Linear forms, 406
Linear mappings, 406
Linear models, see also Nonlinear regression models
Akaike information criterion in, 309
Autocorrelation tests, 129, 132, 176, 207, 209
Confidence regions and, 217, 219, 224
Cox tests in, 293
Davidson–McKinnon tests in, 301
Fisher tests, 72, 169, 180, 305
Homoscedasticity tests in, 208, 209, 211
Inequality constraints and, 237, 245, 261, 271, 272
Information matrix test in, 172
Lagged endogenous variables in, 129, 176
Likelihood ratio tests in, 97, 100
Multiple tests in, 179
Omitted variables tests, 206, 290
Pseudo true values, 278, 279
Residuals, 100
Schwarz criterion in, 311
Score tests in, 97, 99, 206
Selection of regressors in, 288
Specification tests in, 166
Student tests, 57, 180
Wald tests in, 96, 99
Linear systems of equations, 441
Loss
Function, 9
Posterior, 13
Posterior average, 313

Matrix, see also Partitioned matrix, Symmetric matrix
Determinant, 410, 417
Diagonalization, 412, 429
Full-column rank, 411
Identity, 207
Nonsingular, 410
Operations, 409
Orthogonal, 423

Representation of a linear mapping, 408, 411, 423
Square root of, 427
Trace, 413, 440
Transposition, 410, 415
Triangularization, 412
Zero, 409

Maximum likelihood estimation, see also M-estimation
Approximate constrained, 94
Conditional, 83
Consistency, 391
Constrained, 86, 96, 108
GMM estimation and, 164
Inequality constraints and, 243
Pseudo, 151, 170, 278, 383
Quasi generalized pseudo, 153
Mean-value theorem, 400
Measure
Counting, 444
Dirac, 452
Discrete, 446
Integration with respect to a, 446
Lebesgue, 445, 447
Median, 450
M-estimation, see also Extremum estimation
Inequality constraints and, 243
Sequence of local alternatives and, 367
Mill's ratio, 483
Minimax test, 16, 39
Models, see also Disequilibrium models, Exponential models, Linear models, Nonlinear regression models, Panel data model, Simultaneous equations models, Qualitative variable models
Conditional, 34
Identified, 2, 70
Induced, 12, 49
Invariant statistical, 70
Nesting, 234
Nonidentified, 2, 6, 299
Nonparametric, 4
Parametric, 2
Parametric conditional, 163
Second-order, 150
Model selection criteria
Akaike criterion, 307
Bayesian, 312
Linear models and, 304
Schwarz criterion, 311
Moment generating function, see Laplace transform
Monotone likelihood ratio, 28, 38
Multiple tests, see also Sequential tests
Bonferroni procedure, 202
Different models and, 205
Independence of F tests, 181
Linear models and, 179
Scheffé procedure, 203

Newton–Raphson algorithm, 342, 377
Neyman α-structure, 48
Neyman principle, 14, 21
Neyman tests, see also $C(\alpha)$ tests
Definition, 23
Exponential models and, 26
Optimal, 24
Score tests and, 32
Sufficient statistics and, 25
Neyman–Pearson theorem, 22, 67
Noncentrality parameter, 487, 492, 496
Nonlinear regression models
Autocorrelation tests, 117
Davidson–McKinnon tests in, 301
Fisher information matrix in, 101
Homoscedasticity tests in, 116, 131
Lagged endogenous variables in, 118
Omitted-variable tests, 114, 117, 131
Score tests in, 102, 116, 117
Semiparametric efficient estimation in, 377, 380
Semiparametric information bound in, 374
Sequential tests in, 102
Wald tests in, 102
Nonnested hypotheses tests, see also Multiple tests
Comprehensive models and, 297
Cox tests, 293
Generalized Score tests, 287, 292
Generalized Wald tests, 283, 288
Linear models and, 288
Qualitative variable models, 290
Using a nesting model, 297
Nuisance parameters, 43, 47

Omitted-variable tests, 114, 117, 122, 131, 206, 290
Optimal tests, see also Uniformly most powerful tests
Bayesian, 13, 21
Neyman, 24, 26
Nonexistence, 10

Nonrandomized, 25, 26
Power of 25
Overidentifying restrictions tests, 159, 175

Panel data model, 263
Partitioned matrix
Definition, 414
Determinant, 419
Inverse, 417
Pitman approach, 351, 362
Pivotal function
Asymptotic, 225
Definition, 214
Posterior
Distribution, 13, 228
Loss, 13
Probability ratio, 14
Power function
Approximate, 352
Asymptotic, 369
Asymptotic properties, 346
Definition, 9
Optimal tests and, 25
Predictive distribution, 13
Prior
Diffuse, 229
Distribution, 21, 228
Probability, 15
Probability distribution, see also Measure
Absolutely continuous, 447, 459
Bernoulli, 113, 230, 452
Binomial, 30, 452
Cauchy, 35, 69, 453
Chi-square, 487, 492, 514
Conditional, 458, 459
Definition, 443
Discrete, 444
Exponential, 453, 503
Fisher, 495, 516
Gamma, 30, 454, 488
Geometric, 452
Hypergeometric, 65
Joint, 458
Logistic, 454
Marginal, 451, 459
Multinominal, 55, 63, 109, 112
Multivariate normal, 477
Normal, 26, 30, 46, 56, 78, 79, 454, 475, 487, 512, 513
Pareto, 454
Poisson, 27, 30, 78, 452, 488
Student, 492, 515
Truncated normal, 483
Uniform, 37, 452, 453
Probability measure, see Probability distribution
Probability space, 443
Probit model, 121, 132
Projections
Definition, 431
Matrix representations, 433
Orthogonal, 431, 434, 436, 439, 440
Pseudo model, 301, 303
Pseudo true value
Asymptotic, 278, 279
Finite samples (in), 277, 278, 288
P test, 303
Pythagora theorem, 422

Qualitative variable models, 121, 132, 290
Quandt comprehensive model, 298
Quadratic forms
Definition, 420
Gaussian vector (of), 490
Positive definite, 421
Positive semidefinite, 421

Random experiments, 443
Random variables
Conditional independence of, 473
Correlation matrix, 456
Definition, 448
Gaussian, 477, 490
Independence of, 470, 480
Median, 450
Minimum of two normal, 486
Probability distribution of, 448
Standard deviation, 450
Rank
Full-column, 411
Mapping (of), 407
Matrix (of), 411, 440
Rational lags, 156, 175
Residuals, see also Generalized residuals
Diagram, 115
Least squares, 100
Null hypothesis and, 120
Risk
Bayesian, 13, 21
Diagram, 17
Functions, 8
Randomized decision rule and, 9
Second-order approximate quadratic, 336

Root, see Common root tests

Scalar product, 421
Schwarz criterion, 311
Schwarz inequality, 421, 456, 469
Scheffé procedure, 203
Score
  Definition, 68
  General, 69
  Latest exponential model and, 119
Score tests, see also General Asymptotic tests
  Augmented regression and, 114, 290
  Asymptotic critical region, 33
  Asymptotically equivalent tests, 89, 91, 93, 106, 107, 108, 123, 126, 185, 191, 193
  $C(\alpha)$ tests and, 123
  Consistency, 89, 107
  Definition, 32, 87
  Finite sample distribution, 99
  Generalized residuals and, 119
  Hausman tests and, 126
  Independence property, 94, 98, 109, 194
  Likelihood ratio tests and, 91, 100, 107, 185
  Locally uniformly most powerful tests and, 31
  Multiple hypotheses and, 205
  Nonnested hyopotheses and, 287, 292
  Null hypothesis in explicit form and, 106
  Null hypothesis in implicit form and 86, 184
  Principle, 81
  Regression models and, 97, 99, 102, 116, 117, 206
  R^2 and, 90, 100, 103, 115, 116, 117, 118
  Two-sided, 68
  Wald tests and, 89, 100, 106, 185, 362
Second-order asymptotic bias, 337, 340
Selection of regressors, 288
Semiparametric efficiency, 370
Semiparametric information bounds, 372, 374, 376
Semiparametric models, 371
Sequence of local alternative, 348, 362, 366
Sequential tests
  Consistency, 197
  Independence property, 195
  Nested hypotheses in explicit form, 198
  Nested hypotheses in implicit form, 195
  Significance level, 196
Set
  Negligible, 386
  Universally negligible, 387, 389
σ-additivity, 443
σ-algebra, 443
Significance level
  Asymptotic, 84
  Definition, 14
  Limit, 355
  Minimal, 352
  Multiple tests and, 182
  Sequential tests and, 196
Significance tests, 1, 202
Simultaneous equations models, 157, 159, 167, 175
Size functions, 8
Slope
  Approximate, 356, 357
  Exact, 356
Specification tests
  Asymptotic least squares and, 153
  Definition, 202
  Exogeneity tests, 167
  Generalized methods of moments and, 160
  Hausman, 166
  Information matrix test, 170
  Parametric conditional model and, 163
Square root of a matrix, 247
Stability conditions, 236
Standard deviation, 450
Statistic
  Ancillary, 51, 52, 60, 214
  Canonical, 56
  Complete, 48, 211
  Invariant, 71, 72
  Sufficient, 12, 25, 48, 211
Stepwise procedure
  Ascending, 182
  Descending, 182
Stochastic processes
  Definition, 496
  Gaussian, 497
  Poisson, 502
  Stationary, 159, 497, 499, 507
Student tests
  Definition, 59

Fisher tests and, 77
Optimality, 60, 61, 77
Subspaces
Definition, 405
Direct sum of, 407
Orthogonal, 422, 435
Supplementary, 407
Superefficiency, 334
Surjective, 406
Symmetric matrix
Definition, 411
Diagonalization, 429
Ordered spectrum, 425
Ordering, 424
Positive definite, 423
Positive semidefinite, 235, 423, 427

Testing problem
Identified, 3
Invariant, 70
Tests, see also Invariant tests, Likelihood based tests, Nonnested hypotheses tests, Optimal tests, Specification tests, Unbiased tests
Admissable, 11, 21
Autocorrelation, 117, 129, 132, 207, 209
Bayesian, 22
Choice of, 12
Chow, 62
Confidence region and, 222
Consistent, 84
Definition, 6
Exogeneity, 167
Fisher, 72, 180, 305
Function, 7
Homscedasticity, 116, 131, 208, 209, 211
Minimax, 16, 39
Neyman, 23
Nonrandomized, 7, 8, 25, 26
One-sided hypothesis, 28, 30, 49
Ordering, 9
Overidentifying restrictions, 159, 175
Randomized, 7, 8, 28
Ranking, 9
Significance, 1, 202
Student, 59, 77
Two-sided hypothesis, 35, 43, 51
Thresholds, 483, 484
Transition probability, 458

Unbiasedness Principle, 14
Unbiased tests, see also α-similar tests, Uniformly most powerful and unbiased tests
Definition, 41
Locally, 66
unbiased confidence region and, 223
Uniformly most powerful tests, see also α-similar tests, Uniformly most powerful and unbiased tests
Definition, 14
Exponential models and, 26, 30, 36
Invariant and, 71, 75
Locally, 31, 65
Monotone likelihood ratio and, 28
One-sided hypothesis and, 28, 49
Simple hypothesis against simple hypothesis, 17
Two-sided hypothesis and 35, 43, 52
Uniformly most powerful and unbiased tests
Confidence region and, 223
Contingency tables and, 63
Definition, 42
Exponential models and, 43, 49, 52, 54, 77
Locally, 65
Nuisance parameters (in the absence of), 43
One-sided hypothesis and, 49
Scalar linear hypothesis and, 54
Two-sided hypothesis and, 52

Variance
Analysis of, 464, 469
Definition, 450
Residual, 468
Vector space, 405

Wald tests, see also General asymptotic tests
Asymptotically equivalent tests, 89, 91, 93, 107, 108, 123, 126, 185, 191, 193
$C(\alpha)$ tests and, 123
Chi-square tests and, 109
Consistency, 84, 104
Finite sample distribution, 99
Hausman tests and, 126
Independence property, 94, 98, 109, 194
Likelihood ratio tests and, 99, 100, 107, 185
Nonnested hypotheses and, 283, 288

Null hypothesis in explicit form and, 103
Null hypothesis in implicit form and, 84, 184
Principle, 81
Regression models and, 96, 99, 102
Scheffé procedure and, 203
Score tests and, 89, 100, 106, 185, 362
White noise, 497