Monte Carlo Test Methods in Econometrics [*]

Jean-Marie Dufour [†] and Lynda Khalaf [‡]

October 1998
This version: February 1999
Compiled: April 12, 2001, 12:48am

A shorter version of this paper was published in: *Companion to Theoretical Econometrics*, edited by Badi Baltagi, Blackwell, Oxford, U.K., 2001, Chapter 23, 494-519.

[*] This work was supported by the Bank of Canada and by grants from the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Canada Council for the Arts (Killam Fellowship), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds FCAR (Government of Québec). This paper was also partly written at the University of Bristol (Department of Economics) while the first author was Benjamin Meaker Visiting Professor.

[†] Centre de recherche et développement en économique (C.R.D.E.), Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Département de sciences économiques, Université de Montréal. Mailing address: C.R.D.E, Université de Montréal, C.P. 6128 succursale Centre Ville, Montréal, Québec, Canada H3C 3J7. TEL: (514) 343 2400; FAX: (514) 343 5831; e-mail: jean.marie.dufour@umontreal.ca.
Web page: http://www.fas.umontreal.ca/SCECO/Dufour.

[‡] GREEN, Université Laval, Département d'économqiue, Pavillon J.-A. de Sève, Québec, Québec G1K 7P4, Canada. Tel: 1(418) 656-2131-2409; Fax: 1 (418) 656-7412; e-mail: lynda.khalaf@ecn.ulaval.ca

Contents

List of Figures

1. Introduction

During the last 20 years, computer-based simulation methods have revolutionized the way we approach statistical analysis. This has been made possible by the rapid development of increasingly quick and inexpensive computers. Important innovations in this field include the bootstrap methods for improving standard asymptotic approximations [for reviews, see Efron (1982), Efron and Tibshirani (1993), Hall (1992), Jeong and Maddala (1993), Vinod (1993), Shao and Tu (1995), Davison and Hinkley (1997), and Horowitz (1997)] and techniques where estimators and forecasts are obtained from criteria evaluated by simulation [see Mariano and Brown (1993), Hajivassiliou (1993), Keane (1993), Gouriéroux and Monfort (1996) and Gallant and Tauchen (1996)]. An area of statistical analysis where such techniques can make an important difference is hypothesis testing which raises often difficult distributional problems especially in view of determining appropriate critical values.

This paper has two major objectives. First, we review some basic notions on hypothesis testing from a finite-sample perspective, emphasizing in particular the specific role of hypothesis testing in statistical analysis, the distinction between the level and the size of a test, the notions of exact and conservative tests, as well as randomized and non-randomized procedures. Second, we present a relatively informal overview of the possibilities of Monte Carlo test techniques, whose original idea originates in the early work Dwass (1957), Barnard (1963) and Birnbaum (1974), in econometrics. This technique has the great attraction of providing provably *exact* (randomized) tests based on any statistic whose finite sample distribution may be intractable but can be simulated. Further, the validity of the tests so obtained does not depend on the number of replications employed (which can be small). These features may be contrasted with the bootstrap, which only provides asymptotically justified (although hopefully improved) large-sample approximations.

In our presentation, we will try to address the fundamental issues that will allow the practitioners to use Monte Carlo *test* techniques. The emphasis will be on concepts rather than technical detail, and the exposition aims at being intuitive. The ideas will be illustrated using practical econometric problems. Examples include: specification tests in linear regressions contexts (normality, independence, heteroskedasticity and conditional heteroskedasticity), non-linear hypotheses in univariate and SURE models, tests on structural parameters in instrumental regressions, tests on long-run multipliers in dynamic models, long-run identification constraints in VAR models, confidence intervals for ratio of coefficients in discrete choice models. More precisely, we will discuss the following themes.

In Section 2, we identify the important statistical issues motivating this econometric methodology, as an alternative to standard procedures. The issues raised have their roots in practical test problems and may be outlined as follows.

- An *Exact* Test Strategy: What Is It, and Why Should We Care?
- The *Nuisance-Parameter* Problem: What Does It Mean to Practitioners?
- Understanding the *Size/Level* Control Problem.
- *Pivotal* and *Boundedly-Pivotal* Test Criteria: Why Is this Property Important?

- Identification and Near non-identification: A Challenging Setting.
- Non-Nested hypothesis.

Further, the relevance and severity of the problem will be demonstrated using simulation studies and/or empirical examples.

Sections 3 and 4 describe the Monte Carlo (MC) test method along with various econometric applications of it. Among other things, the procedure is compared and contrasted with the bootstrap. Whereas bootstrap tests are asymptotically are asymptotically valid (as both the numbers of observations and simulated samples go to ∞), a formal demonstration is provided to emphasize the size control property of MC tests. Monte Carlo tests are typically discussed in parametric contexts. Extensions to non-parametric problems are also discussed. The theory is applied to a broad spectrum of examples that illustrate the usefulness of the procedure. We conclude with a word of caution on inference problems that cannot be are solved by simulation. For convenience, the concepts and themes covered may be outlined as follows.

- MC Tests Based on Pivotal Statistics: An Exact Randomized Test Procedure
- MC Tests in the Presence of Nuisance Parameters
  - ▶ Local MC p-value
  - ▶ Bounds MC p-value
  - ▶ Maximized MC p-value
- MC Tests versus the Bootstrap
  - ▶ Fundamental Differences/Similarities
  - ▶ The Number of Simulated Samples: Theory and Guidelines
  - ▶ MC tests or Bartlett Corrections?MC tests: Breakthrough Improvements and ”Success Stories”
  - ▶ The Intractable Null Distributions Problem (*e.g.* tests for normality, Uniform Linear hypothesis in Multi-Equation models, tests for ARCH)
  - ▶ The Case of Unidentified Nuisance Parameters (test for structural jumps, test for ARCH-M)
  - ▶ Non-Nested Tests
  - ▶ Induced Tests: The “Combination” Problem
  - ▶ Non-Standard, Possibly Unbounded Confidence Intervals
- MC Tests May Fail: Where and Why? A Word of Caution.

We conclude in Section 5.

2. Statistical issues: a practical approach to core questions

> *Is it more serious to convict an innocent man or to acquit a guilty? That will depend on the consequences of the error; is the punishment death or fine; what is the danger to the community of released criminals* [...]. *From the point of view of mathematical theory all that we can do is to show how the risk of errors may be controlled and minimized.*
> [Neyman and Pearson (1933), quoted by McCloskey and Ziliak (1996)]

In order to understand the theory of hypothesis testing, it is necessary to analyze the intuitive reasoning underlying the classical [Neyman-Pearson] criteria for constructing a "good" test.

The hypothesis testing problem is often presented as one of deciding between two hypotheses: the hypothesis of interest (the *null* H_0) and its complement (the *alternative* H_A). For the purpose of the exposition, consider a test problem pertaining to a *parametric* model $(\mathcal{Y}, \mathsf{P}_\theta)$, *i.e.* the case where the data generating process [DGP] is determined up to a **finite** number of unknown real parameters $\theta \in \Theta$. Θ refers to the parameter space, $\mathcal{Y}$ is the sample space, P_θ is the family of probability distributions on $\mathcal{Y}$, and let Y denote the observations. Furthermore, let Θ_0 refer to the subspace of Θ compatible with H_0.

A statistical test partitions the sample space into two subsets: a set consistent with H_0 (the region of acceptance), and its complement whose elements are viewed as inconsistent with H_0 (the region of rejection, or the **critical region**). This may be translated into a decision rule based on a *test statistic* $S(Y)$: the rejection region is defined as the numerical values of the test statistic for which the null will be rejected. In general, a test is completely defined by its critical region. When the observed value of the test statistic falls in the critical region, the result is said to be **significant**.

Associated with any rejection decision are two possible outcomes:

- the test may reject H_0 when H_0 is true; this is called **type I error**;
- the test may reject H_0 when H_0 is false; the probability of this is called **the power** of the test.

A test whose error probability is as small as possible is clearly desirable. Furthermore, the ability of the test to detect departures from the null (the probability of the right answer) is also of concern. The central issue is, and will remain, how to reconcile both desirable yet somewhat conflicting characteristics. As might be expected, a test which minimizes risk and at the same time maximizes power is almost impossible to devise. The following discussion outlines the elements of the **classical** approach to the problem that is due to Neyman and Pearson.

The starting point in the test construction process is to modify the fundamental questions. Instead of addressing the test problem as one of choosing between H_0 and H_A, the question is restated as follows:

"Is there enough evidence to reject the null hypothesis?"

Accordingly, the potential test outcomes are:

- "the test rejects H_0 : the sample provides convincing evidence in favor of H_A";

- “the test fails to reject H_0 : the sample evidence is insufficient to disprove H_0”.

This - weaker and rather ”**conservative**” - formulation of the problem is the one that has received the most attention from statisticians and has led to the most commonly used test strategy: set the type I error probability to a small pre-assigned value α and then, subject to this constraint, find criteria with ”good” power. In other words, restrict focus to procedures which “**control**” the risk of falsely rejecting the null and among those, select the ones which optimize the probability of rejecting a false null.

> Minimal risk actually means controlled risk; maximum power cannot be achieved arbitrarily but is subject to ”risk” control.

Then the concept of “a most powerful test” becomes meaningless![1] We are rather led to consider “ most powerful $\alpha\%$ tests”, a notation which suggests limiting considerations to tests of some preassigned error probability.

The concepts underlying the optimality criteria just stated have often been compared to the principles of court justice. Under “standard” judicial systems, the “alleged criminal” is **presumed innocent until proven guilty**. Convictions require **proving guilt beyond reasonable doubt**. In terms of hypotheses tests, failing to reject the null is analogous to failing to convict a defendant, in which case the person on trial will either be innocent or the evidence was not decisive enough to overturn the initial presumption of innocence.

2.1. Significance level, size and power: basic definitions

In our discussion so far, we have avoided the terms *significance level* and *test size.* Instead, we have focused on the intuitive interpretation of type I error control. We move now to the formal translation of this concept into mathematical statements.

Let R refer to the rejection region of a test. The associated function describing the type I error probability may be defined as

$$\alpha(\theta) = \mathsf{P}_\theta(R)$$

where θ satisfies the null. In this framework, an $\alpha\%$ test implies

$$\alpha(\theta) \leq \alpha \tag{2.1}$$

for all $\theta \in \Theta_0$.This may be restated as follows

$$\sup_{\theta\in\Theta_0} \alpha(\theta) \leq \alpha\ . \tag{2.2}$$

α is called the “significance **level**” (notice the $\leq$ sign) and $\sup[\alpha(\theta)], \theta \in \Theta_0$ is called the **size** of the test. In other words, a test of size α implies (notice the $=$ sign)

$$\sup_{\theta\in\Theta_0} \alpha(\theta) = \alpha\ . \tag{2.3}$$

[1] If we adopt such a criterion, we might as well ignore the data and merely decide, apriori, to reject the null.

At this stage, it is worth emphasizing that although a size correct procedure is preferred, the classical test theory requires primarily the control of the significance level. The fact that in some cases, it is practically impossible to determine a critical point so that size is exactly α does not matter. We just chose an α-level critical region whose size is as nearly as possible α. In other words, level control is required yet size control - if possible - is desirable.

The mathematical implications of size or level control can be quite complicated. Indeed, in the probability statements (2.2) and (2.3), the sample size is explicitly taken into consideration. In addition, the underlying supremum, in many cases, is far from trivial.

> When we talk about an **exact test,** it must be understood that attention is restricted to level-correct critical regions, where (2.2) is evaluated in terms of a fixed sample size, for all values of the intervening parameter compatible with the null.

We now turn to the concept of power. Formally, the power of a test may be defined as

$$\mathsf{P}_\theta(R)\,, \quad \theta \notin \Theta_0 \tag{2.4}$$

where R is determined by considerations of the level of the test. For a given α, a test which satisfies the condition

$$\begin{array}{ll} \mathsf{P}_\theta(R) \leq \alpha\,, & \text{for all } \theta \in \Theta_0\,, \\ \mathsf{P}_\theta(R) \text{ is maximum}, & \text{for all } \theta \notin \Theta_0\,, \end{array}$$

is said to be *uniformly most powerful at level* α, *i.e.* most powerful against all possible DGP's that do not satisfy the null. Unfortunately, for most problems of practical interest, it is the exception rather than a rule that such a test exists. Usually, attention is restricted to alternatives of particular interest. To solve this difficulty, we may impose further reasonable restrictions on the class of tests considered. For instance, we expect the power of a test to exceed its size, for all $\theta \notin \Theta_0$. Otherwise, the test is more liable to accept H_A when it is false than when it is true. A test satisfying

$$\begin{array}{ll} \mathsf{P}_\theta(R) \leq \alpha\,, & \text{for all } \theta \in \Theta_0\,, \\ \mathsf{P}_\theta(R) \;\geq \alpha\,, & \text{for all } \theta \notin \Theta_0\,, \end{array}$$

is said to be an **unbiased α-level test.** Generally speaking, we also expect that larger samples provide more information,*i.e.* more convincing evidence in favor or against H_0. If the sample size is **sufficiently large**, we expect to reject a false null with probability close to one. This defines the **consistency** property. A test is consistent (for a certain class of alternatives) if the power $\to 1$ as the sample size $\to \infty$. Furthermore, for certain problems, it seems natural to demand that the test be **invariant** under certain reparameterizations (transformations of θ).

Another natural restriction on the class of tests is worth considering. Suppose it is possible to find a rejection region such that $\mathsf{P}_\theta(R)$ does not depend on θ under the null. This clearly simplifies the analytical problems associated with (2.3). A test with the property

$$\mathsf{P}_\theta(R) = \alpha, \quad \text{for all } \theta \in \Theta_0\,,$$

is said to be **similar.**

2.2. Large sample tests

It is useful at this stage to restate the definitions introduced above in terms of the familiar framework of test statistics. Without loss of generality, suppose the critical region takes the form

$$S(Y) \geq c$$

where $S(Y)$ is a test statistic. To obtain an α-level test, c must be chosen so that

$$\sup_{\theta \in \Theta_0} \mathsf{P}_\theta[S(Y) \geq c] \leq \alpha \,. \tag{2.5}$$

The test will have size α if and only if

$$\sup_{\theta \in \Theta_0} \mathsf{P}_\theta[S(Y) \geq c] = \alpha \,. \tag{2.6}$$

To solve for c in (2.5) or (2.6), it is necessary to extract the finite sample distribution of $S(Y)$ when the null is true. Typically, $S(Y)$ is a complicated function of the observations and the statistical problem involved is often intractable. More importantly, it is evident from the definitions (2.5), (2.6), (2.2) and (2.3) that in many cases of practical interest, the distribution of $S(Y)$ may be different for different parameter values. Unless of course the null fixes the values of θ (*i.e.* Θ_0 is a point); such hypotheses are called *simple hypotheses*. Most hypothesis encountered in practice are **composite** i.e. the set Θ_0 has more than one element. The null may uniquely define some parameters, but almost invariably, some other parameters are not restricted to a point-set. In the context of composite hypotheses, some unknown parameters may appear in the distribution of $S(Y)$. Such parameters, for obvious reasons, are called **nuisance parameters.** Consequently, in carrying out an exact test, one may encounter two difficulties. First is the problem of extracting the analytic form of the distribution of $S(Y)$. Second is the problem of maximizing the rejection probabilities over the relevant nuisance parameter space. The first difficulty is easily removed in the context of Monte Carlo testing. We draw attention to the latter problem, the importance of which is not fully recognized in econometric practice.

As it is often the case, a reasonable solution for both problems exists when one is dealing with large samples. Whereas the null distribution of $S(Y)$ may be complicated and/or may involve unknown parameters, the asymptotic null distribution of the commonly used test statistic has a known form and is nuisance-parameter-free. Then the critical point may conveniently be obtained using asymptotic arguments. The term **approximate critical point** is more appropriate here, since we are dealing with asymptotic levels: the choice of the cut-off point which yields an approximate size-α test may be very different from the exact critical values. To illustrate these issues, we next present two commonly used test criteria: the likelihood ratio (LR) and the Wald test statistics.

2.2.1. The likelihood ratio test

Consider the case where both H_0 and H_A are simple hypotheses, i.e. the sample space has only two elements θ_0 and θ_1, and the hypothesis takes the form

$$\begin{aligned} H_0 :& \quad \theta = \theta_0 \, , \\ H_A :& \quad \theta = \theta_1 \, . \end{aligned}$$

Suppose the associated probability distributions are defined by density functions $f(y;\theta_0)$ and $f(y;\theta_1)$. It can be shown (using the Neyman-Pearson fundamental Lemma) that any critical region of the form

$$R^* = \{ \ y : \quad f(y;\theta_1)/f(y;\theta_0) \geq k \ \}$$

where k is chosen such that

$$\mathsf{P}_{\theta_0}(R^*) = \alpha,$$

is uniformly most powerful at level α. Note that the term "uniformly" is unnecessary here because H_A reduces to one point.

The above result provides the basis of the LR principle and justifies the following generalization to composite hypotheses. As emphasized earlier, in this case $f(y;\theta)$ is not uniquely determined under the null nor under the alternative, and the Neyman-Pearson fundamental Lemma cannot be readily applied. We arrive at a reasonable procedure by the following modification. Given a sample, determine its "best chance" if H_A is true and its "best chance" if H_0 is true. Large enough values of the ratio of these best chances are interpreted as evidence against H_0. The best chance of a sample translates into the well known maximum likelihood (ML) principle which yields the following rejection region:

$$R^{**} = \left\{ \ y : \ \left(\sup_{\theta \notin \Theta_0} f(y;\theta) \right) / \left(\sup_{\theta \in \Theta_0} f(y;\theta) \right) \geq k \ \right\}$$

where k is determined by the condition

$$\sup_{\theta \in \Theta_0} \mathsf{P}_{\theta_0}(R^{**}) = \alpha \, .$$

The LR principle provides a formal method for constructing critical regions. Note that no asymptotic argument has been used so far. Although the LR principle is not justified on asymptotic grounds, asymptotic theory conveniently provides approximate critical points. Indeed, it can be shown, under general regularity conditions, that the asymptotic distribution of

$$-2 \ln \left[\frac{\sup_{\theta \notin \Theta_0} f(y;\theta)}{\sup_{\theta \in \Theta_0} f(y;\theta)} \right]$$

is χ^2 with known degrees of freedom and the test is asymptotically most powerful.

The LR test can be slightly modified to improve the χ^2 approximation. Examples include the Bartlett correction. Bartlett corrections involve rescaling the test statistic by a suitable constant obtained such that the mean of the scaled statistic equals that of the approximating distribution to a given order [Bartlett (1937), Barndorff-Nielsen and Blaesild (1986)]. The formulae for the relevant constant must be solved on a case by case basis.

We conclude this preliminary discussion on LR test with the following note. The large-sample distributional results underlying the χ^2 approximation are valid if Θ_0 is a subset of the space compatible with H_A, defined by r non-redundant restrictions (**nested** hypotheses), in which case the constraints in H_0 may be reduced by reparameterization to the form

$$h_j(\theta) = 0\,,\ i = j,\ \ldots\ ,\ r. \tag{2.7}$$

In fact r will determine the degrees of freedom of the statistic's limiting distribution. It is usual in this context to refer to the restricted and unrestricted models. We will emphasize that the application of the Monte Carlo test technique to LR criteria allows one to obtain valid tests with non-nested hypotheses. This is indeed a very useful property.

2.2.2. The Wald test

In the context of hypotheses of the form (2.7), the Wald criterion is based on the vector $h_j(\widehat{\theta})$, where $\widehat{\theta}$ is the ML estimate. The null is rejected if $h(\widehat{\theta})$ is far enough from zero, where $h(\widehat{\theta})$ denotes the $r \times 1$ vector with elements $h_j(\widehat{\theta})$. As it is well known, under general regularity conditions, we have:

$$\sqrt{n}(\widehat{\theta} - \theta) \overset{asy}{\sim} N(0, B_\theta^{-1}) \tag{2.8}$$

where B_θ is the information matrix. Using a Taylor expansion, we can write

$$h(\widehat{\theta}) \simeq h(\theta) + H_\theta'(\widehat{\theta} - \theta)$$

where

$$H_\theta = \frac{\partial h_j(\theta)}{\partial \theta_j}\,.$$

Under the null hypothesis $h(\theta) = 0$, then

$$h(\widehat{\theta}) \simeq H_\theta'(\widehat{\theta} - \theta)\,.$$

It follows that

$$\sqrt{n}h(\widehat{\theta}) \overset{asy}{\sim} N(0, H_\theta' B_\theta^{-1} H_\theta)\,. \tag{2.9}$$

This may be used to derive a region of "proximity to zero" around $h(\widehat{\theta})$ and yields the following critical region for testing (2.7):

$$\left\{\ y:\ \ nh'(\widehat{\theta})\left[H_{\widehat{\theta}}' B_{\widehat{\theta}}^{-1} H_{\widehat{\theta}}\right]^{-1} h(\widehat{\theta}) \geq k\ \right\}. \tag{2.10}$$

The test statistic so defined is asymptotically $\chi^2(r)$ under the null. At this stage, it is worth emphasizing that the principle underlying the Wald test is fundamentally asymptotic, in the sense that (2.8), (2.9) and (2.10) are justified only on asymptotic grounds. In the case of the LR test, we resorted to asymptotic theory at the stage of cut-off points computations. In terms of the Wald test, there is no reason why the region of "proximity to zero"[2] implied by (2.9) should be preferred (or even relevant) from the finite sample perspective.

2.2.3. How large is large?

For sufficiently large sample sizes, the standard asymptotic approximations are expected to work well. The question is, and will remain, **how large is large**? To illustrate this issue, we next consider several examples involving commonly used econometric methods. We will demonstrate, by simulation, that asymptotic procedures may yield highly unreliable decisions, with samples of sizes empirically relevant. The problem, and our main point, is that **finite sample accuracy is not merely a small sample problem.**

2.3. Econometric applications: Monte Carlo studies

2.3.1. Instrumental regressions [3]

Consider the *limited-information* (LI) structural regression model:

$$\begin{aligned} y &= Y\beta + X_1\gamma_1 + u = Z\delta + u\,, \\ Y &= X_1\Pi_1 + X_2\Pi_2 + V\,, \end{aligned} \tag{2.11}$$

where Y and X_1 are $n \times m$ and $n \times k$ matrices which respectively contain the observations on the included endogenous and exogenous variables, $Z = [Y, X_1]$, $\delta = (\beta', \gamma_1')'$and X_2 refers to the excluded exogenous variables. If more than m variables are excluded from the structural equation, the system is said to be *over-identified.* The associated LI reduced form is:

$$\left[\begin{array}{cc} y & Y \end{array}\right] = X\Pi + \left[\begin{array}{cc} v & V \end{array}\right],\ \Pi = \left[\begin{array}{cc} \pi_1 & \Pi_1 \\ \pi_2 & \Pi_2 \end{array}\right], \tag{2.12}$$

$$\pi_1 = \Pi_1\beta + \gamma_1\,,\ \ \pi_2 = \Pi_2\beta\,. \tag{2.13}$$

The necessary and sufficient condition for identification follows from the relation $\pi_2 = \Pi_2\beta$. Indeed β is recoverable if and only if

$$rank(\Pi_2) = m\,. \tag{2.14}$$

To test the general linear hypothesis $R\delta = r$, the well-known IV analogue of the Wald test is frequently applied on grounds of computational ease. For instance, consider the two-stage least squares (2SLS) estimator

$$\hat{\delta}_i = [Z_i'P(P'P)^{-1}P'Z_i]^{-1}Z_i'P(P'P)^{-1}P'y_i \tag{2.15}$$

[2] Using standard notation, this region is the familiar $\widehat{\theta} \pm t_{\alpha/2}\,\mathrm{SE}(\widehat{\theta})$ in the scalar parameter case.

[3] This section is based on the results in Dufour and Khalaf (1998b).

where P is the following matrix of instruments $P = [X_i\,,\ X(X'X)^{-1}X'Y_i]$. Application of the Wald principle yields the following criterion

$$\tau_w = \frac{1}{s^2}(r - R\hat{\delta}_i)' - [R'(Z_i'P(P'P)^{-1}P'Z_i)^{-1}R]\ (r - R\hat{\delta}_i) \tag{2.16}$$

where

$$s^2 = \frac{1}{n}(y_i - Z_i\hat{\delta}_i)'(y_i - Z_i\hat{\delta}_i)'. \tag{2.17}$$

Under usual regularity conditions and imposing identification, τ_w is distributed like a $\chi^2(q)$ variable.

Bartlett (1948) and Anderson and Rubin (1949, AR) suggested an exact test that can be applied only if the null takes the form $\beta = \beta^0$. The idea behind the test is quite simple. Define $y^* = y - Y\beta^0$. Under the null, the model can be written as $y^* = X_1\gamma_1 + u$. On the other hand, if the hypothesis is not true, y^* will be a linear function of all the exogenous variables. Thus, the null may be assessed by the F test that the coefficient of the "excluded" regressors is zero in the regression of y^* on all the exogenous variables.

We first consider a simple experiment based on the work of Nelson and Startz (1990a, 1990b) and Staiger and Stock (1997). The model considered is a special case of (2.11) with two endogenous variables ($p = 2$) and $k = 1$ exogenous variables. The structural equation includes only the endogenous variable. The restrictions tested were of the form

$$H_{01} : \beta = \beta^0\,. \tag{2.18}$$

The sample sizes were set to $n = 25, 100$. The exogenous regressors were independently drawn from the normal distribution, with mean zero and unit variance. These were drawn only once. The errors were generated according to a multinormal distribution with mean zero and covariance

$$\Sigma = \begin{bmatrix} 1 & .95 \\ .95 & 1 \end{bmatrix}. \tag{2.19}$$

The other coefficients were:

$$\beta = \beta^0 = 0\,;\ \Pi_2 = 1\,,\ .9\,,\ .7\,,\ .5\,,\ .2\,,\ .1\,,\ .05\,,\ .01. \tag{2.20}$$

In this case, the 2SLS-based test corresponds to the standard t-test [see Nelson and Startz (1990b) for the relevant formulae]. 1000 replications were performed, and Figures 1.1 - 1.3 report the estimated 97.5 percentile for the two-tailed 2SLS t-test for the significance of β. Figures 1.4 - 1.6 present the estimated probabilities of type I error [$P(type\,I\,error)$] associated with the latter test.

In this context, the identification condition reduces to $\Pi_2 \neq 0$. This condition can be tested using a standard F test in the first stage regression. The problem is more complicated when the structural equation includes more than one endogenous variable. To examine this case, we consider the following simulation experiment.[4] The model is (2.11) with three endogenous variables ($m = 2$)

[4] See Dufour and Khalaf (1998b) for a detailed discussion of this experiment.

Figure 1.1: IV-based t-test critical points, sample size = 25

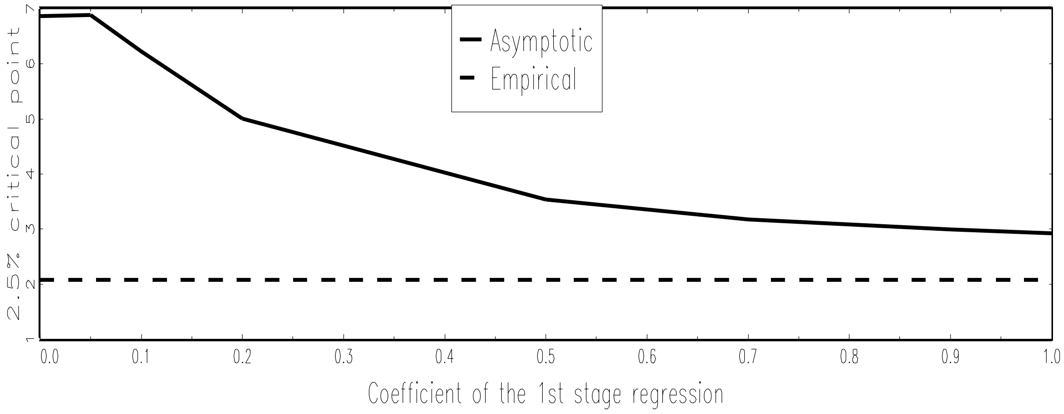


Figure 1.2: IV-based t-test critical points, sample size = 100

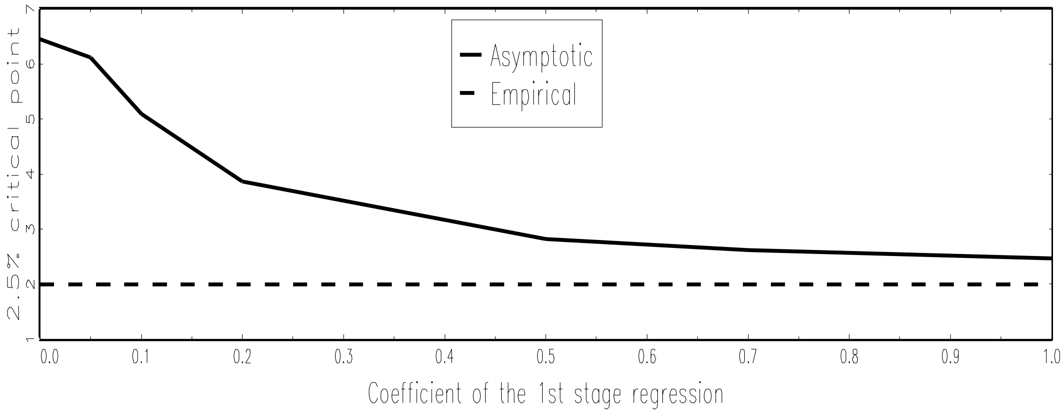


Figure 1.3: IV-based t-test critical points, sample size = 250

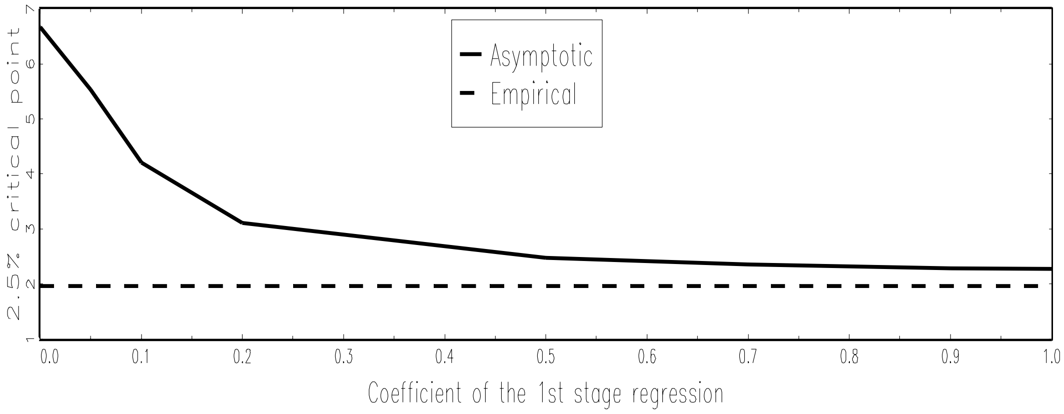


Figure 1.4: IV-based t-test / Anderson-Rubin F test, sample size = 25

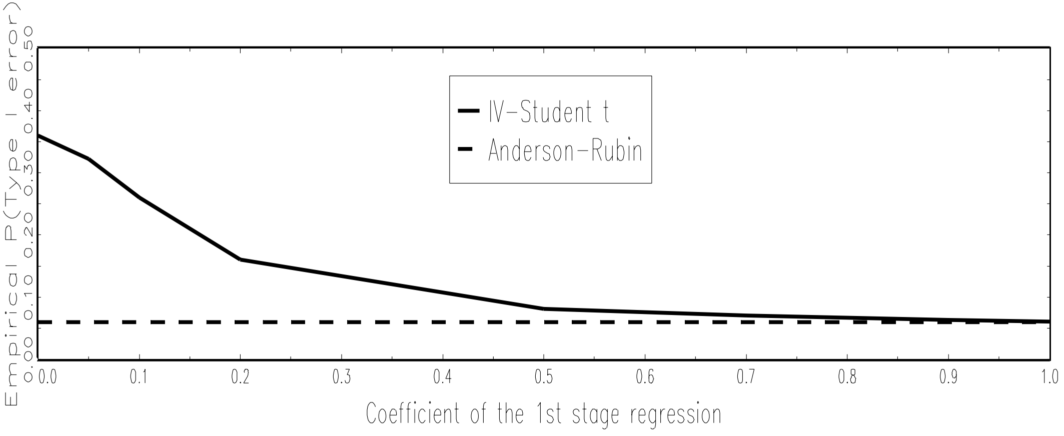


Figure 1.5: IV-based t-test / Anderson-Rubin F test, sample size = 100

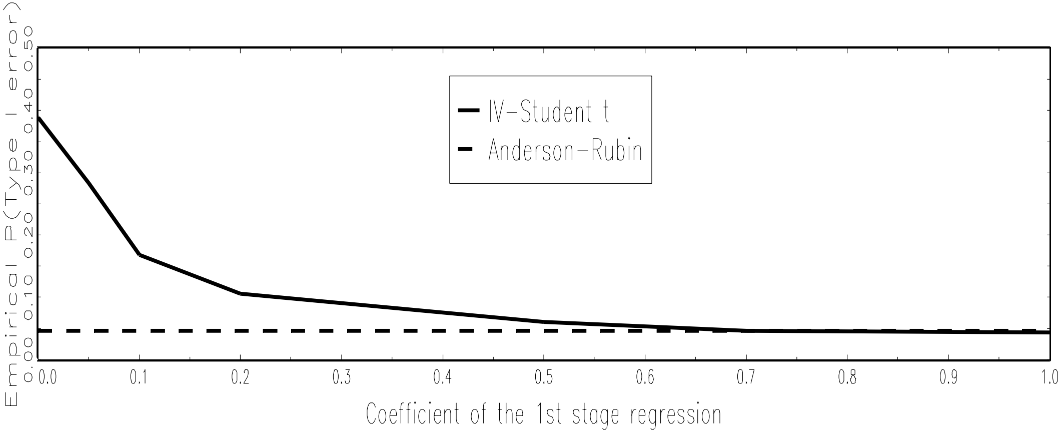
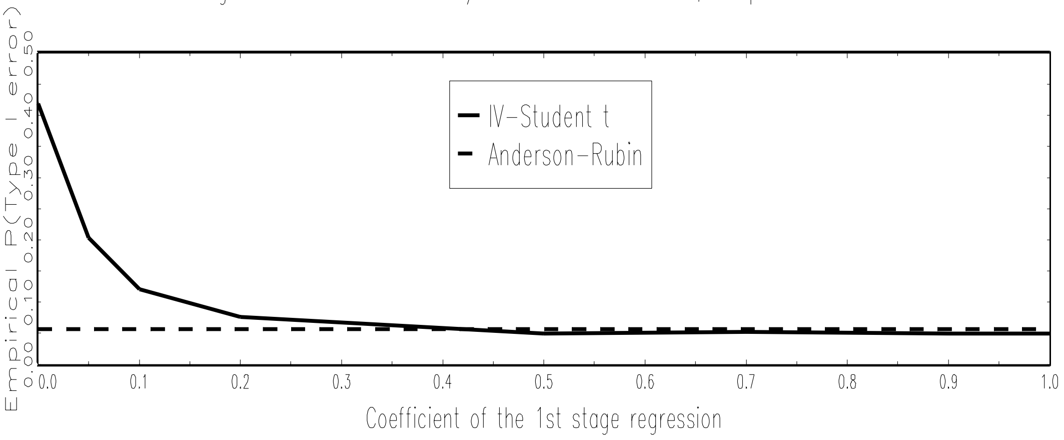


Figure 1.6: IV-based t-test / Anderson-Rubin F test, sample size = 250



and $k = 6$ exogenous variables. In all cases, the structural equation includes only one exogenous variable, the constant regressor. The restrictions tested were of the form

$$H_{01} : \beta = \beta^0 \, , \tag{2.21}$$

$$H_{02} : \beta_1 = \beta_1^0 \, , \tag{2.22}$$

where $\beta = (\beta_1, \ \beta_2)'$. The sample sizes were set to $n = 25$, 100. The exogenous regressors were independently drawn from the normal distribution, with means zero and unit variances. These were drawn only once. The errors were generated according to a multinormal distribution with mean zero and covariance

$$\Sigma = \begin{bmatrix} 1 & .95 & -.95 \\ .95 & 1 & -1.91 \\ -.95 & -1.91 & 12 \end{bmatrix} . \tag{2.23}$$

The other coefficients were:

$$\gamma_1 = 1 \, , \ \beta_1 = 10 \, , \ \beta_2 = -1.5 \, , \ \Pi_1 = (1.5, 2)' \, , \ \Pi_2 = \begin{bmatrix} \Pi^\iota \\ O_{(k-3,2)} \end{bmatrix} \tag{2.24}$$

with

$$\Pi^\iota = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{2.25}$$

The identification problem becomes mores serious as the determinant of $\Pi_2'\Pi_2$ gets closer to zero. In view of this, we also consider:

$$\begin{aligned} \Pi^\iota &= \begin{bmatrix} 2 & 1.999 \\ 1.999 & 2 \end{bmatrix} , \quad \Pi^\iota = \begin{bmatrix} .5 & .499 \\ .499 & .5 \end{bmatrix} , \\ \Pi^\iota &= \begin{bmatrix} .1 & .099 \\ .099 & .1 \end{bmatrix} , \quad \Pi^\iota = \begin{bmatrix} .01 & .009 \\ .009 & .01 \end{bmatrix} . \end{aligned} \tag{2.26}$$

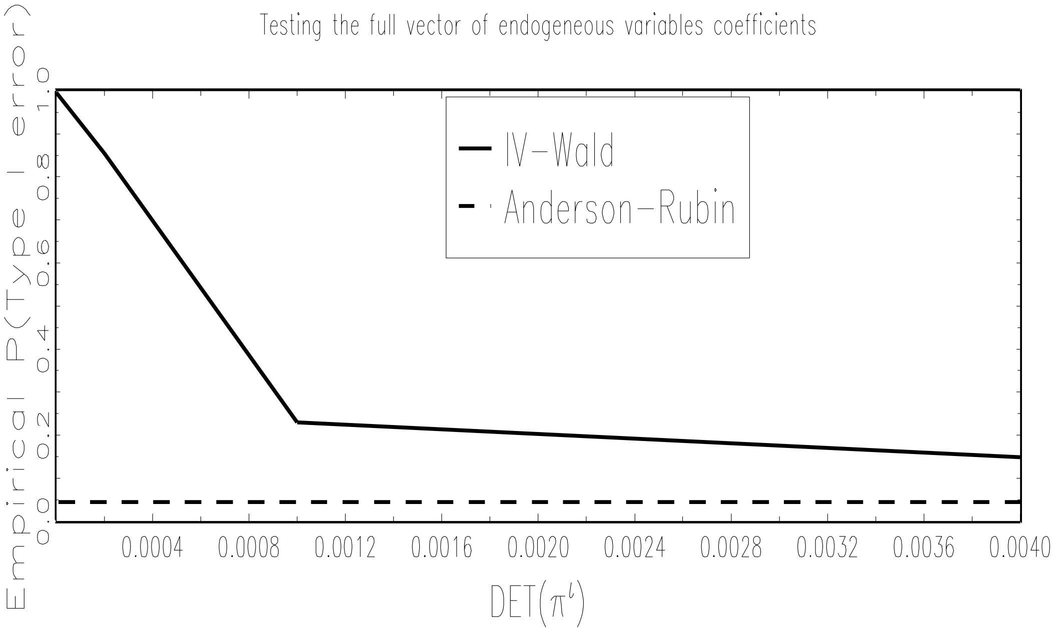
For further reference, the model corresponding to (2.25) is called the “identified” case and the model corresponding to (2.26) the ”near-unidentified” case. The Wald statistics based on 2SLS were calculated as defined above. The standard asymptotic χ^2 critical value was adopted and 1000 replications were performed. Figures 2.1- 2.4 summarize our findings for the near-unidentified case. The results for the identified model are as follows.

Table 1 : Empirical $P(Type\ I\ error)$, identified model

Sample size	Wald(H_{01})	Anderson-Rubin(H_{01})	Wald(H_{02})
25	14.8	4.4	14.2
100	8.4	4.3	8.1

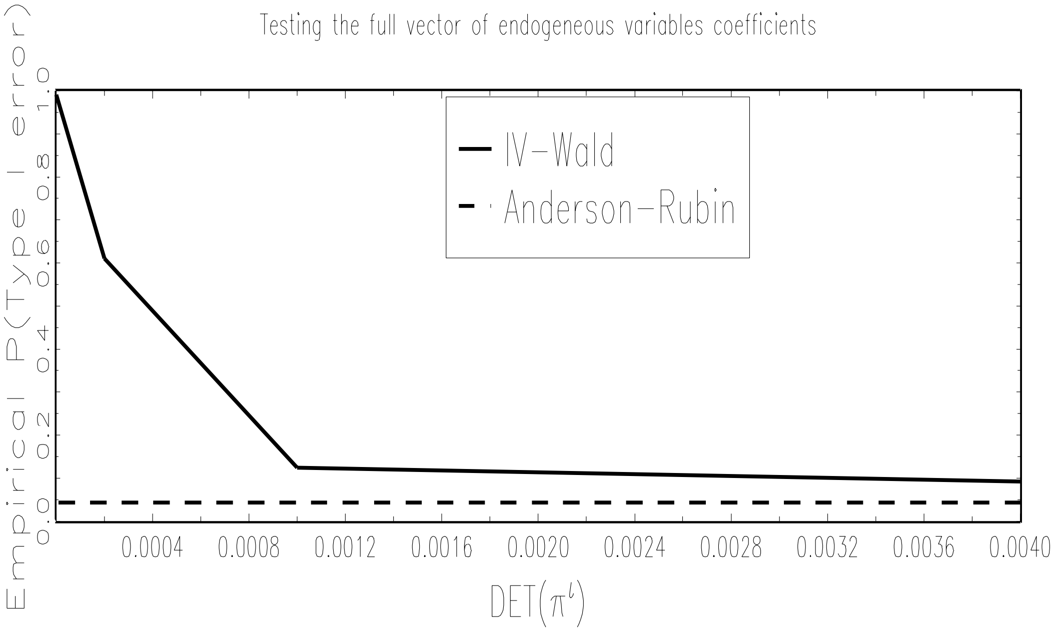
It is evident from both experiments that IV-based Wald tests perform very poorly in terms of size control. Identification problems severely distort the test sizes. While the evidence of size distortions is notable even in identified models, the problem is far more severe in near-unidentified situations.

Figure 2.1: IV-based Wald test / Anderson-Rubin F test, sample size = 25



Testing the full vector of endogeneous variables coefficients

Figure 2.2: IV-based Wald test / Anderson-Rubin F test, sample size = 100



Testing the full vector of endogeneous variables coefficients

Figure 2.3: IV-based Wald test, sample size = 25

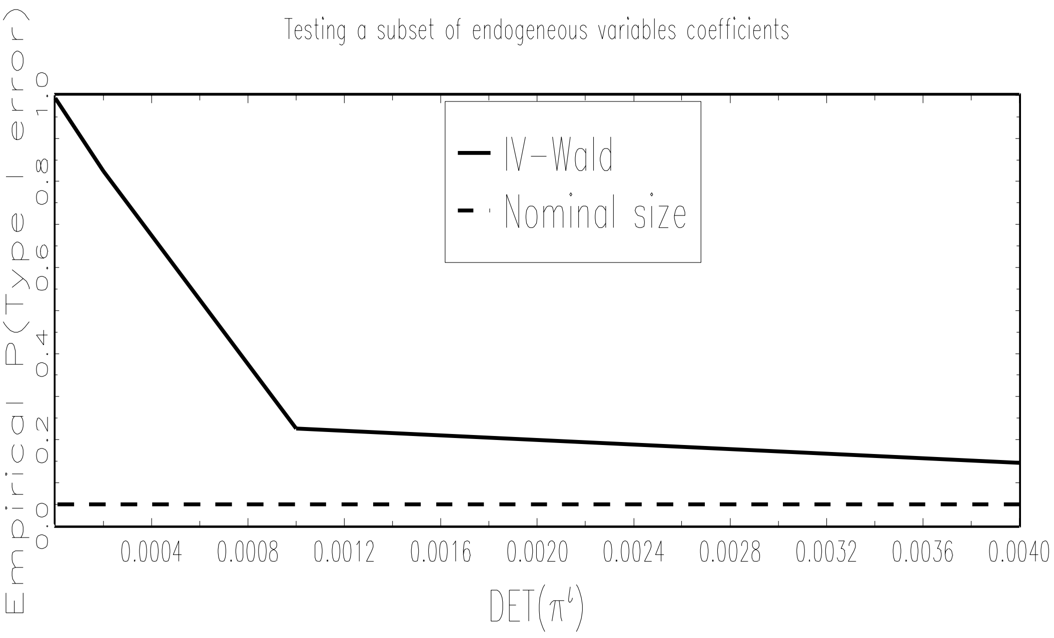
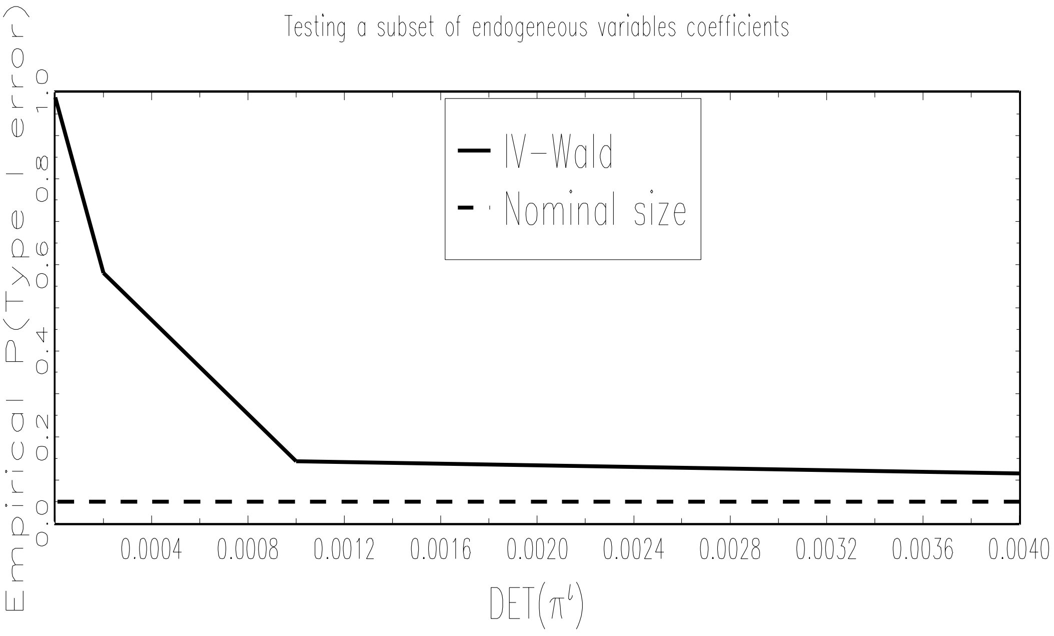

Testing a subset of endogeneous variables coefficients

Figure 2.4: IV-based Wald test, sample size = 100


Testing a subset of endogeneous variables coefficients

More importantly, increasing the sample size does not correct the problem. In this regard, Bound, Jaeger, and Baker (1995) report severe bias problems associated with IV-based estimators, despite very large sample size.

> "*The use of large data sets does not necessarily insulate researchers from quantitatively important finite sample biases.*" [Bound, Jaeger, and Baker (1995)]

In contrast, the Anderson-Rubin test, when available, is immune to such problems: the test is exact, in the sense that the null distribution of the AR criterion does not depend on the parameters controlling identification. Indeed, the AR test statistic follows an $F(m, n-k)$ distribution, regardless of the identification status. The AR test has recently received renewed attention; see, for example, Dufour and Jasiak (1994) and Staiger and Stock (1997). Recall however that the test is not applicable unless the null sets the values of the coefficients of all the endogenous variables.

Despite the recognition of the need for caution in the application of IV-based tests, the standard econometric software packages typically implement IV-based Wald tests. In particular, the t-tests on individual parameters are routinely computed in the context of 2SLS or 3SLS procedures. Unfortunately, the Monte Carlo experiments we have analyzed confirm that IV-based tests realize computational savings at the risk of very poor performance.

2.3.2. Normality tests[5]

We consider normality tests in the context of the linear regression model:

$$Y = X\beta + u \tag{2.27}$$

where $Y = (y_1, \ldots, y_n)'$ is a vector of observations on the dependent variable, X is the matrix of n observations on k regressors, β is a vector of unknown coefficients and u is an n-dimensional vector of $i.i.d$ disturbances. The problem consists in testing:

$$H_0 : f(u) = \varphi(0, \sigma) \; , \; \sigma > 0 \, , \tag{2.28}$$

where $f(u)$ is the unknown probability density function (pdf) and $\varphi(\mu, \sigma)$ is the normal pdf with mean μ and standard deviation σ. In this context, normality tests are typically based on the least-squares residual vector

$$\widehat{u} = y - X\widehat{\beta} = M_X u \tag{2.29}$$

where $\widehat{\beta} = (X'X)^{-1} X'y$ and $M_X = I_n - X(X'X)^{-1} X'$. Let $\widehat{u}_{1n} \leq \widehat{u}_{2n} \leq \ldots \leq \widehat{u}_{nn}$ denote the order statistics of the residuals, and

$$s^2 = (n-k)^{-1} \sum_{i=1}^{n} \widehat{u}_{in}^2 \; , \quad \widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \widehat{u}_{in}^2 \; . \tag{2.30}$$

Here we focus on two tests: the Kolmogorov-Smirnov (KS) test [Kolmogorov (1933), Smirnov (1939)], and the Jarque and Bera (1980, 1987; henceforth JB) test.

[5] This section is based on the results in Dufour, Farhat, Gardiol, and Khalaf (1998).

The KS test is based on a measure of discrepancy between the empirical and hypothesized distributions. The exact and limiting distributions of the test statistic are non-standard and even asymptotic points must be estimated. The statistics are defined as follows:

$$KS = \max\left(D^{+},\ D^{-}\right) \tag{2.31}$$

where $D^{+} = \max_{1\leq i\leq n}\left[(i/n) - \widehat{z}_i\right]$ and $D^{-} = \max_{1\leq i\leq n}\left[\widehat{z}_i - (i-1)/n\right]$,where $\widehat{z}_i = \Phi(\widehat{u}_{in}/s),\ i = 1,\ ...,\ n$, and $\Phi(.)$ denotes the cumulative $N(0,1)$ distribution function. We have used significance points from D'Agostino and Stephens (1986, Table 4.7), although these formally were derived for the location-scale model.

The JB test derives from the recognition that the third and fourth moments of the $N(0,1)$ distribution are equal to 0 and 3 respectively. Hence deviations from normality may be assessed using the sample moments, *i.e.* the coefficients of skewness (Sk) and kurtosis (ku):

$$Sk = n^{-1}\sum_{i=1}^{n}\widehat{u}_{in}^{3}/(\widehat{\sigma}^{2})^{3/2}\,, \tag{2.32}$$

$$Ku = n^{-1}\sum_{i=1}^{n}\widehat{u}_{in}^{4}/(\widehat{\sigma}^{2})^{2}\,. \tag{2.33}$$

The skewness and kurtosis tests may be implemented as two distinct tests [see D'Agostino and Stephens (1986, chapter 6)]. The JB Lagrange multiplier criterion combined both Sk and ku :

$$JB = n\left[\frac{1}{6}(Sk)^{2} + \frac{1}{24}\left(Ku-3\right)^{2}\right]. \tag{2.34}$$

Under the null and appropriate regularity conditions, the JB statistic is asymptotically distributed as $\chi^{2}\left(2\right)$. As is typically the case with the various normality tests, the exact distribution is intractable.

The simulation experiment was performed as follows. The model used was (2.27). For each disturbance distribution, the tests were applied to the residual vector, obtained as $\widehat{u} = M_{x}u$. Hence, there was no need to specify the coefficients vector β. The matrix X included a constant term, k_1 dummy variables, and a set of independent standard normal variates. Formally,

$$X = \left[\ \iota_n \vdots\ X_{(1)} \vdots\ X_{(2)}\ \right],\ X_{(1)} = \left[\begin{array}{l} I_{k_1} \\ Z_{(n-k_1,k_1)} \end{array}\right] \tag{2.35}$$

where $Z_{(i,j)}$ denotes an (i,j) matrix of zeros, $X_{(2)}$ includes $k - k_1 - 1$ regressors drawn as independent and identically distributed $(i.i.d.)$ standard normal. Sample sizes of $n = 25,\ 50,\ 100$ (and 300 in certain cases) were used, k was set as the largest integer less than or equal to $\sqrt{n}$ and $k_1 = 0,\ 2,\ 4,\ ...,\ k$. The disturbances were generated from the standard normal. Figures 4.1 - 4.2 report rejection percentages (from 10000 replications) at the nominal size of 5%. "Design 1", "Design 2" and "Design 3" refer to the following:

$Design\ 1$	$k_1 = 0$
$Design\ 2$	$\begin{pmatrix} k_1 = 2 \\ n = 25 \end{pmatrix}, \begin{pmatrix} k_1 = 4 \\ n > 25 \end{pmatrix}$
$Design\ 3$	$\begin{pmatrix} k_1 = k \\ n \leq 50 \end{pmatrix}, \begin{pmatrix} k_1 = 8 \\ n = 100 \end{pmatrix}, \begin{pmatrix} k_1 = 10 \\ n = 300 \end{pmatrix}$

Our conclusions may be summarized as follows. Although the tests appear adequate when the explanatory variables are generated as standard normal, the sizes of all tests vary substantially from the nominal 5% for all other designs, irrespective of the sample size. More specifically, (i) the KS test consistently overrejects, and (ii) the JB test based on $\widehat{\sigma}$ underrejects when the number of dummy variables relative to normal regressors is small and overreject otherwise.

2.3.3. Confidence sets for parameter ratios in discrete choice models [6]

A common problem in discrete choice models consists in building confidence regions for possibly nonlinear parameter transformations. For instance, in the analysis of random utility travel demand models, one is frequently required to construct a confidence interval (CI) for ratios of parameters. Usually, this is performed using the $delta$-method that generates Wald-type confidence intervals based on consistent parameter and variance/covariance estimates. To fix ideas, consider the simple binary probit model

$$\mathsf{P}(Y_i = 1) = F(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) \, , \; i = 1, \; \dots \; , \; n, \tag{2.36}$$

where $F(.)$ is the cumulative standard normal distribution. The problem is to construct a confidence interval for the ratio β_1/β_2. Given MLE $\widehat{\beta}_1$ and $\widehat{\beta}_2$ and an associated variance/covariance matrix $\widehat{\Sigma}_{12}$, the *delta*-method yields the following $(1-\alpha)$ CI:

$$\left[(\widehat{\beta}_1/\widehat{\beta}_2) \mp z_{\alpha/2}(\widehat{g}'\widehat{\Sigma}_{12}\widehat{g})^{1/2}\right] \tag{2.37}$$

where

$$\widehat{g} = \left[\begin{array}{cc} 1/\widehat{\beta}_2 \, , & \widehat{\beta}_1/\widehat{\beta}_2^2 \end{array} \right]'$$

and $z_{\alpha/2}$ refers to the standard normal deviate. As it is well known, in discrete response models, such CI's are only asymptotically valid. There is substantial evidence that standard asymptotics provide a poor approximation to the sampling distribution of estimators and test statistics in the presence of discrete or limited dependent variables; see, for example Davidson and MacKinnon (1999c, 1999a, 1999b) and Savin and Würtz (1998). These papers focus on linear hypothesis tests, and this certainly casts doubts on the reliability of procedures involving parameter ratios, where the underlying inferential problem is more complicated. In this regard, the analytical results in Dufour (1997) have serious implications for the problem at hand. Indeed, it is shown that standard CI's for ratios of coefficients have confidence levels that may deviate arbitrarily from their nominal level, since they are almost surely bounded. The following simulation experiment based on (2.36) was

[6] This section is based on the results in Bolduc, Dufour, and Khalaf (1998).

Figure 4.1: Kolmogorov–Smirnov residual–based test

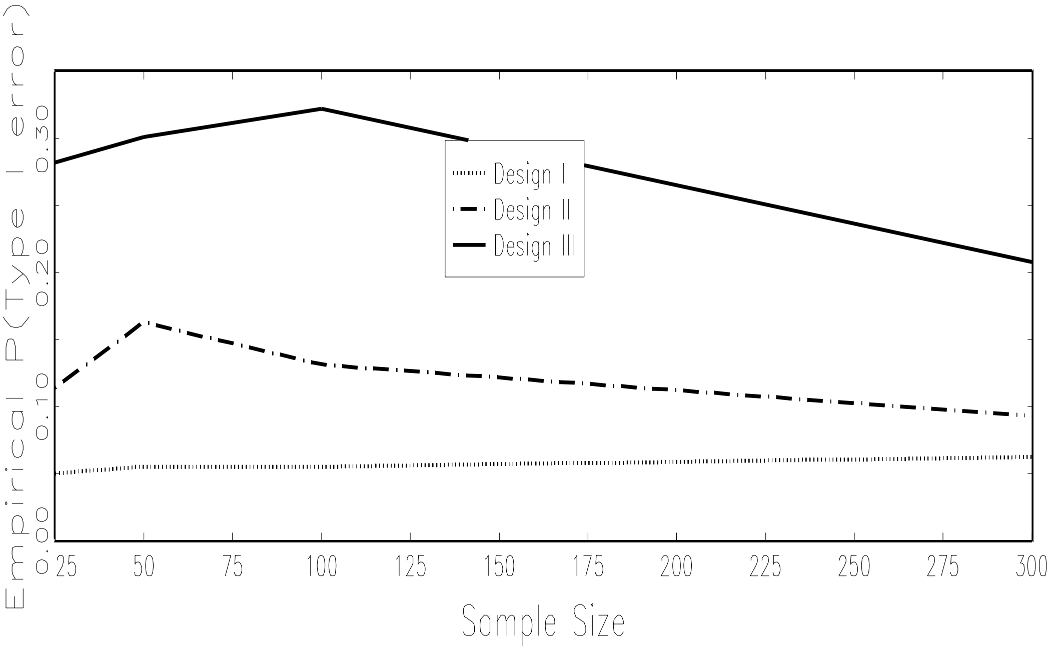
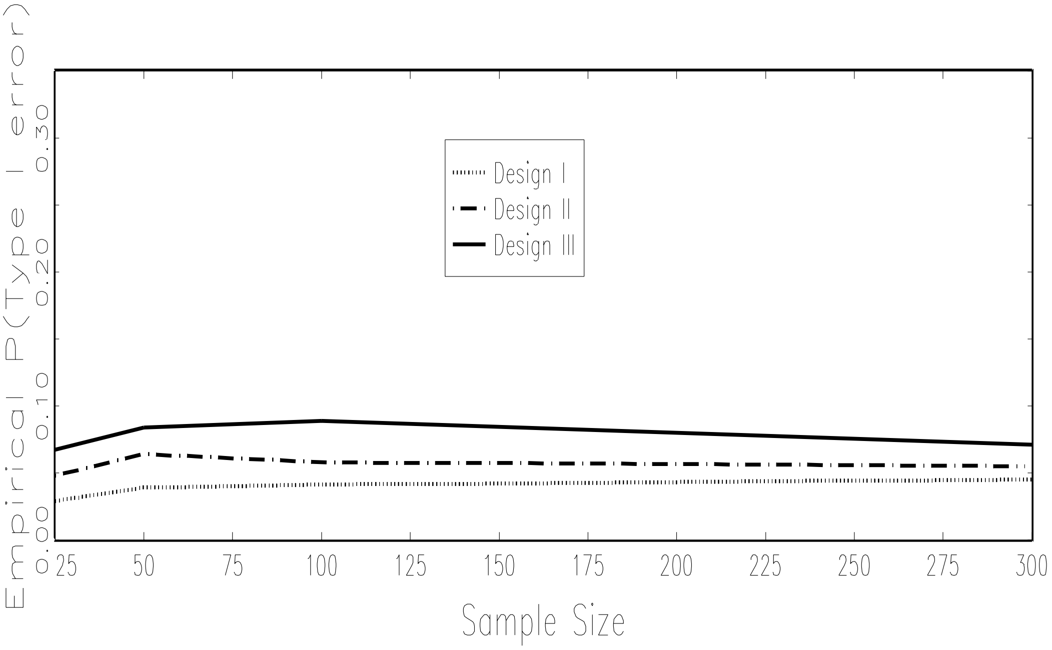


Figure 4.2: Jarques–Bera test



designed to assess the severity of this problem. We considered $n = 100,\ 250$, and set:

$$\beta_0 = 1\,;\ \beta_1 = 1\,;\ \beta_2 = 1\,,\ .5\,,\ .4\,,\ .3\,,\ .2\,,\ .1\,,\ .01\,,\ .001\,.$$

The regressors were drawn as standard normal. As is well known, in this simple "regular" framework, the likelihood function is well behaved and the asymptotic variance-covariance matrix estimate is given by $\widehat{\Sigma} = (X'DX)^{-1}$, where X is the regressor matrix, D is a diagonal $n \times n)$ matrix with entries

$$d_i = \frac{[f(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})]^2}{F(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})\,[1 - F(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})]}$$

and $f(.)$ is the standard normal probability density function. The CI was obtained as in (2.37). The coverage probabilities for a nominal 95% confidence level are reported in Figures 3.1 - 3.2.

These findings illustrate quite clearly that the use of the *delta*-method may lead to unreliable inference. Indeed, it can be seen from Figures 3.1 - 3.2 that the true coverage probabilities are substantially different from 95%, with the coverage sharply deteriorating as β_2 approaches zero. Moreover, increasing the sample size does not seem to solve the problem. The practical implication is that CI obtained as in (2.37) may have zero confidence levels; in other words, the probability that the above CI "misses" the true ratio is practically "one".

2.3.4. Uniform linear hypothesis in multivariate regression models [7]

Multivariate linear regression (MLR) models involve a set of p regression equations with cross-correlated errors. When regressors may differ across equations, the model is known as the Seemingly Unrelated Regression model [SURE, Zellner (1962)]. The MLR model can be expressed as follows:

$$Y = XB + U \tag{2.38}$$

where $Y = [Y_1,\ \ldots\ ,\ Y_p]$ is an $n \times p$ matrix of observations on p dependent variables, X is an $n \times k$ full-column rank matrix of fixed regressors, $B = [\beta_1,\ \ldots\ ,\ \beta_p]$ is a $k \times p$ matrix of unknown coefficients and $U = [U_1,\ \ldots\ ,\ U_p] = [\widetilde{U}_1,\ \ldots\ ,\ \widetilde{U}_n]'$ is an $n \times p$ matrix of random disturbances with covariance matrix Σ where $\det(\Sigma) \neq 0$. An alternative representation of the model is

$$Y_{ij} = \alpha_j + \sum_{k=1}^{p} \beta_{jk} X_{ik}\ ,\ i = 1\ ,\ \ldots\ ,\ n,\ \ j = 1\ ,\ \ldots\ ,\ p\,. \tag{2.39}$$

Uniform Linear (UL) constraints take the special form

$$H_0 : RBC = D \tag{2.40}$$

[7] This experiment is taken from Dufour and Khalaf (1998a).

Figure 3.1: Ratio of Parameters

Probit model, sample size=100

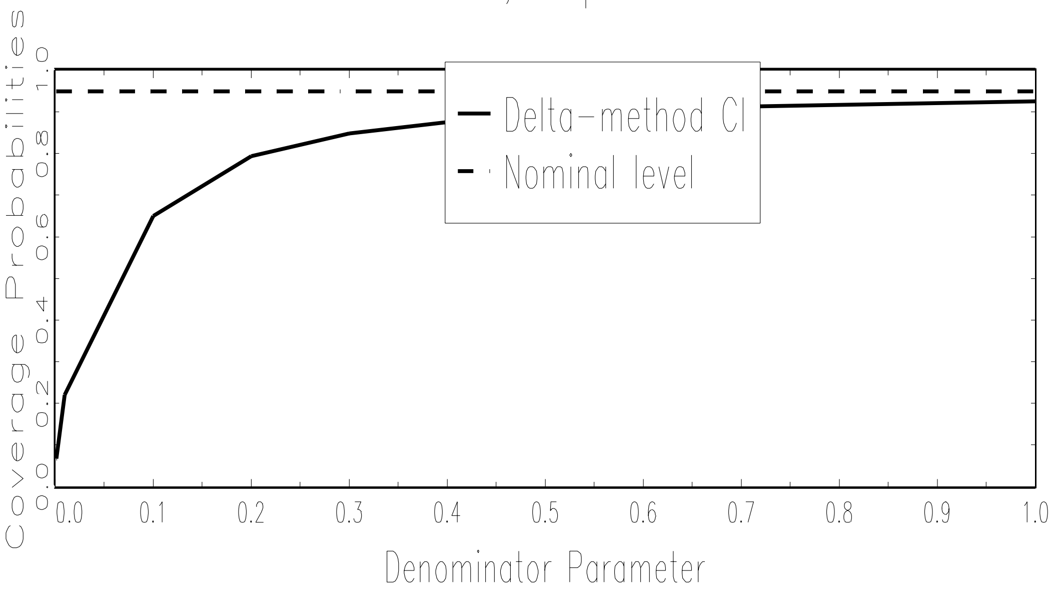
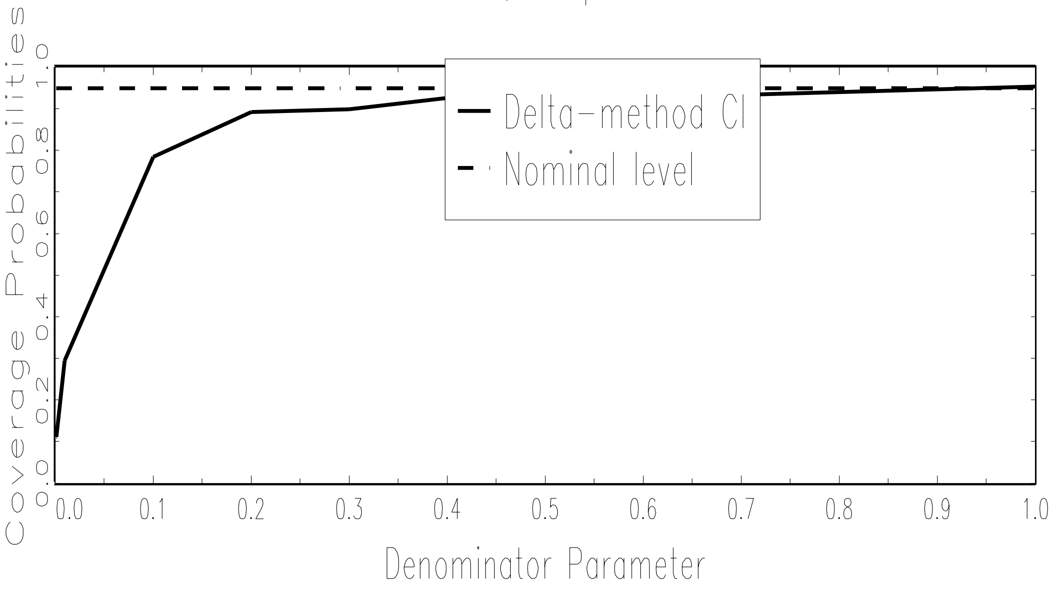


Figure 3.2: Ratio of Parameters

Probit model, sample size=250



where R is a known $r \times k$ matrix of rank $r \leq k$, C is a known $p \times c$ matrix of rank $c \leq p$, and D is a known $r \times c$ matrix. An example is the case where the same hypothesis is tested for all equations

$$H_{01} : R\beta_i = \delta_i \, , \; i = 1, \, \ldots \, , \, p \, , \tag{2.41}$$

which corresponds to $C = I_p$. In this context, the likelihood ratio (LR) criterion is:

$$LR = n \; \ln(\Lambda) \, , \; \Lambda = |\widehat{U}_0'\widehat{U}_0| / |\widehat{U}'\widehat{U}| \; ; \tag{2.42}$$

where $\widehat{U}_0'\widehat{U}_0$ and $\widehat{U}'\widehat{U}$ are respectively the constrained and unconstrained SSE matrices.

Stewart (1997) discusses several econometric applications where the problem can be stated in terms of (2.40). A prominent example includes the multivariate test of the capital asset pricing model CAPM. Let $R_{jt}, \; j = 1, \, \ldots \, , \, p,$ be security returns for period $t, \; t = 1, \, ... \, , T.$ If it is assumed that a riskless asset R_F exists, then efficiency can be tested based on the following MLR-based capital asset pricing model (CAPM)

$$R_{jt} - R_{Ft} = \alpha_j + \beta_j (R_{Mt} - R_{Ft}) + \epsilon_{jt} \, , \; j = 1 \, , \, \ldots \, , \, p \, , \quad t = 1 \, , \, \ldots \, , \, T \, ,$$

where R_{Mt} are the returns on the market benchmark. The hypothesis of efficiency implies that the intercepts α_j are jointly equal to zero, *i.e.*

$$\alpha_j = 0 \, , \; j = 1 \, , \, \ldots \, , \, p. \tag{2.43}$$

The latter hypothesis is a special case of (2.40) with $C = I_p$ and R is the $1 \times p$ vector $(1, \, 0, \, ... \, , 0)$. Another example concerns demand analysis. It can be shown [see for example Berndt (1991, Chapter 9)] that the translog demand specification yields a model of the form (2.39). In this context, the hypothesis of linear homogeneity takes the form

$$H_0 : \sum_{k=1}^{p} \beta_{jk} = 0 \, , \; j = 1 \, , \, ... \, , \, p \, .$$

We consider a simulation experiment modelled after the latter application, with

$$p = 5, \, 7, \, 8, \quad n = 20 \, , \; 25 \, , \; 40 \, , \; 50 \, , \; 100 \, .$$

The regressors are independently drawn from the normal distribution; the errors are independently generated as $i.i.d. \; N(0, \Sigma)$ with $\Sigma = GG'$ and the elements of G drawn (once) from a normal distribution. The coefficients for all experiments are available from Dufour and Khalaf (1998a). The statistics examined are the relevant LR criteria defined by (2.42) and the Bartlett-corrected LR test [Attfield (1995, section 3.3)]. The results are summarized in Figures 5.1 - 5.2. We report the tests empirical size, based on a nominal size of 5% and 1000 replications. It is evident that the asymptotic LR test overreject substantially. Second, the Bartlett correction, though providing some improvement, fails in larger systems. In this regard, it is worth noting that Attfield (1995, section 3.3) had conducted a similar Monte Carlo study to demonstrate the effectiveness of Bartlett adjustments in this framework, however the example analyzed was restricted to a two-equations model.

To conclude this section, it is worth noting that an exact test is available for the special cases

$$H_0 : RBC = D\,, \quad \min(r, c) \leq 2\ .$$

Indeed, Laitinen (1978) in the context of the tests of demand homogeneity and Gibbons, Ross, and Shanken (1989), for the problem of testing the CAPM efficiency hypothesis, independently show that a transformation of the relevant LR criterion has an exact F distribution given normality of asset returns.[8]

2.3.5. Testing for ARCH and GARCH [9]

Consider the linear model

$$y_t = x_t'\beta + u_t\,,\ t = 1,\ \ldots\ ,\ T, \tag{2.44}$$

where $x_t = (1,\ x_{t2},\ \ldots\ ,\ x_{tk})'$, β is a $k \times 1$ vector of unknown coefficients and the error terms u_t are $i.i.d.$ with mean 0 and variance σ_t^2. We are concerned with the problem of testing the null hypothesis

$$H_0 : \sigma_t^2 = \sigma^2\,,\ t = 1,\ \ldots\ ,\ T, \tag{2.45}$$

against the $ARCH(m)$, or $GARCH(q, m)$ alternative. The standard test [Engle (1982), Lee (1991)] may be obtained as T times the R^2 from regression of $\widehat{u}_t^2$ on a constant and m lags of $\widehat{u}_t$, where $\widehat{u}_t$ denotes the OLS residuals estimate. Under H_0, the Engle test statistic follows a $\chi^2(m-1)$ in large samples. We investigate the performance of this test using the following design. The regressors are drawn as standard normal, β is a $k \times 1$ vector of ones, $\sigma^2 = 1, T = 25,\ 50,\ 100$, and k is set as the largest integer less than or equal to $\sqrt{T}$. The results are shown in Figure 6. It is evident that the test is undersized.

2.4. Econometric applications: discussion

In many empirical problems, it is quite possible that the exact null distribution of the relevant test statistic $S(Y)$ is not easy to compute analytically yet it is nuisance-parameter-free (recall the definition of similar tests). In this case $S(Y)$ is called a pivotal statistic, i.e. the null distribution of $S(Y)$ is uniquely determined under the null hypothesis. In such cases, we will show that the MC test easily solves the size control problem, regardless of the distributional complexities involved. The above examples on normality tests, heteroskedasticity tests and the uniform linear hypothesis tests, all involve pivotal statistics.

The problem is more complicated in the presence of nuisance parameters. To conclude this section, we will state a property related to test statistics which will prove to be fundamental in finite sample contexts.[10]

[8] The underlying distributional result is due to Wilks (1932).

[9] This section is based on Bernard, Dufour, Khalaf, and Genest (1998).

[10] For a formal treatment see Dufour (1997).

Figure 5.1: LR-based tests, Ho uniform-Linear
MLR model, 5 Equations

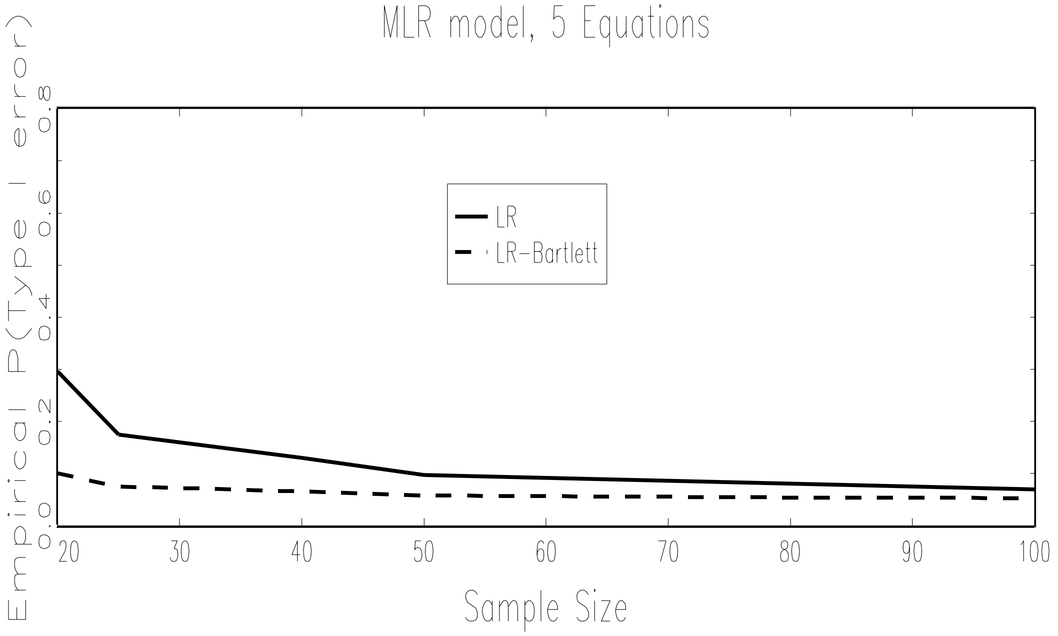


Figure 5.2: LR-based tests, Ho uniform-Linear
MLR model, 8 Equations

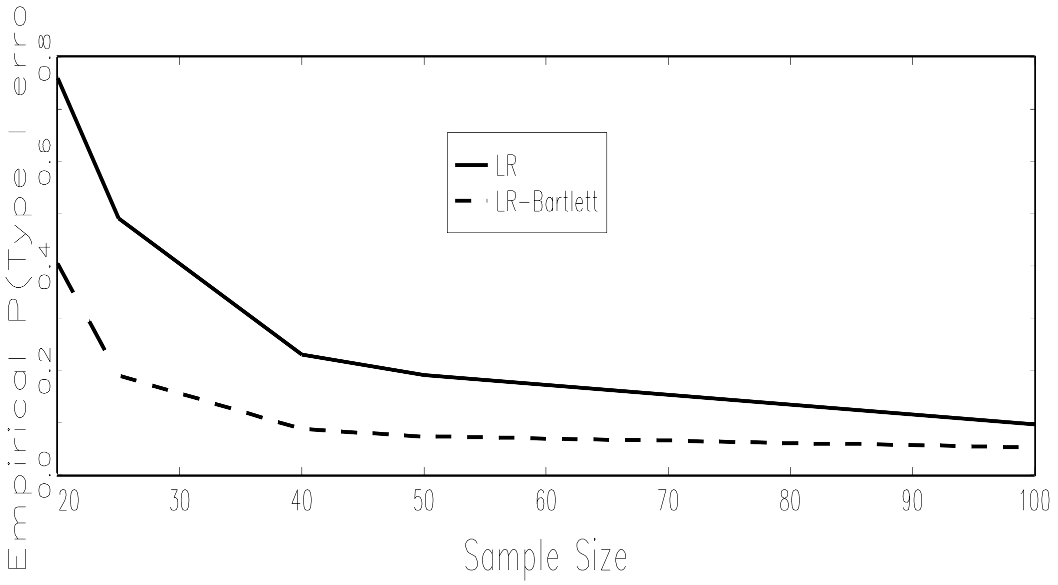
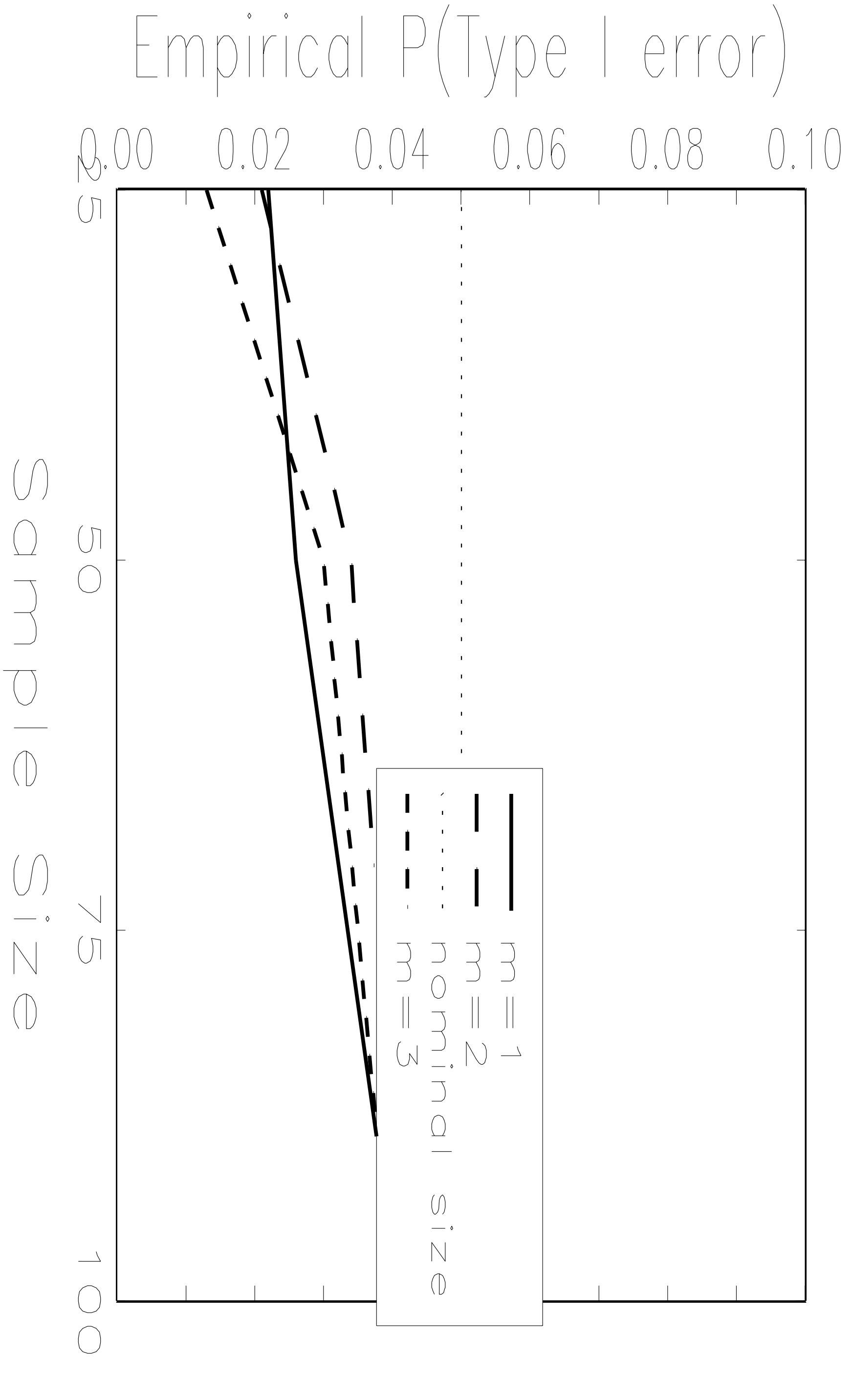


Figure 6: LM tests for ARCH(m) and GARCH(q,m



In the context of a right-tailed test problem, consider a statistic $S(Y)$ whose null distribution depends on nuisance parameters and suppose that it is possible to find another statistic $S^*(Y)$ such that

$$S(Y) \leq S^*(Y), \quad \forall\theta \in \Theta_0\,, \tag{2.46}$$

and $S^*(Y)$ is pivotal under the null. Then $S(Y)$ is said to be **boundedly pivotal.** The implications of this property are as follows. From (2.46), we obtain

$$\mathsf{P}_\theta[S(Y) \geq c] \leq \mathsf{P}[S^*(Y) \geq c]\,, \quad \forall\theta \in \Theta_0\,.$$

Then if we calculate c such that

$$\mathsf{P}[S^*(Y) \geq c] = \alpha\,, \tag{2.47}$$

we solve the level constraint for the test based on $S(Y)$. It is clear that (2.46) and (2.47) imply

$$\mathsf{P}_\theta[S(Y) \geq c] \leq \alpha\,, \quad \forall\theta \in \Theta_0\ .$$

As emphasized earlier, the size control constraint is easier to deal with in the case of $S^*(Y)$ because it is pivotal. Consequently, the maximization problem

$$\sup_{\theta\in\Theta_0} \mathsf{P}_\theta[S(Y) \geq c]$$

has a non-trivial solution (less than 1) in the case of **boundedly pivotal statistics**. If this property fails to hold, the latter optimization problem may admit only the trivial solution, so that it becomes mathematically impossible to control the significance level of the test.

It is tempting to dismiss such considerations assuming they will occur only in "textbook" cases. Yet it can be shown (we will consider this issue in the next section) that similar considerations explain the poor performance of the Wald tests and confidence intervals in examples 1 and 3 above. **These are problems of empirical relevance in econometric practice.** In the next session, we will show that the bootstrap will also fail for such problems!

We close this section with these comments from Phillips (1983):

> "*For the process by which asymptotic machinery works inevitably washes out sensitivities that are present and important in finite samples. Thus generality and robustness in asymptotic theory are achieved at the price of insensitivity with respect to ingredients as the distributional characteristics of a model's random elements and the values of many of its parameters.*"

3. The Monte Carlo test technique: an exact randomized test procedure

> *If there were a machine that could check 10 permutations a second, the job would run something on the order of 1000 years. The point is, then, that an impossible test can be*

made possible, if not always practical. [Dwass (1957)]

The Monte Carlo (MC) test procedure was first proposed by Dwass (1957) in the following context. Consider two independent samples $X_1,\ ...,\ X_m$ and $Y_1,\ ...,\ Y_n$ where the X's are $\overset{iid}{\sim}\ (F(x))$, the Y's are $\overset{iid}{\sim}\ (F(x-\delta))$ and the *cdf* $F(.)$ is continuous. No further distributional assumptions are imposed. To test $H_0 : \delta = 0$, the following procedure may be applied.

- Let
$$z = (X_1,\ ...,\ X_m,\ Y_1,\ ...,\ Y_n)\ ,$$
$$s = \frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i\,.$$

- Obtain all possible $Q = (n+m)!$ permutations of z, $z^{(1)},\ ...,\ z^{(Q)}$, and calculate the associated "permuted analogues" of s
$$s^{(j)} = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} z_i^{(j)} - \tfrac{1}{n}\sum_{i=m+1}^{m+n} z_i^{(j)}\ ,\quad j = 1,\ \ldots\ ,\ Q\,.$$

- Let r denote the number of $s^{(j)}$'s for which $s \leq s^{(j)}$. Reject the null (*e.g.* against $H_A : \delta > 0$) if $r \leq k$, where k is a predetermined integer.

It is easy to see that
$$\mathsf{P}(r \leq k) = k/Q$$
under the null because the X's and the Y's are exchangeable. In other words, the test just described is exactly of size k/Q.

The procedure is intuitively appealing, yet there are $(n+m)!$ permutations to examine. To circumvent this problem, Dwass (1957) proposed to apply the same principle to a sample of P permutations $\widetilde{s}^{(1)}$, ..., $\widetilde{s}^{(P)}$, **in a way that will preserve the size of the test**. The modified test may be applied as follows:

- Let $\widetilde{r}$ denote the number of $\widetilde{s}^{(j)}$'s for which $s \leq \widetilde{s}^{(j)}$. Reject the null (against $\delta > 0$) if $\widetilde{r} \leq d$, where d is chosen such that
$$\frac{d+1}{P+1} = \frac{k}{Q}.$$

Dwass formally shows that with this choice for d, the size of modified test is exactly k/Q = the size of the test based on all permutations. This means that if a 5% permutation test is desired, and 99 permutations are manageable, then $d+1$ should be set to 5. The latter decision rule may be restated as follows: reject the null if the rank of s in the series
$$s,\ \ \widetilde{s}^{(1)},\ ...,\ \widetilde{s}^{(P)}$$
is less than or equal to 5.

Since each $\widetilde{s}^{(j)}$ is "weighted" by the probability that it is sampled from all possible permutations, the modification due to Dwass yields a randomized test procedure.

The principles underlying the MC test procedure are highly related to the randomized permutation test just described. Indeed, this technique is based on the above test strategy where the sample of permutations is replaced by **simulated samples**. Note Barnard (1963) proposed later a similar idea.

3.1. Monte Carlo tests based on pivotal statistics

In the following, we briefly outline the MC test methodology as it applies to the pivotal statistic context and a right tailed test; for a more detailed discussion, see Dufour (1995) and Dufour and Kiviet (1998).

Let S_0 denote the observed test statistic S, where S is the adopted test criterion. We assume S has a unique continuous distribution under the null hypothesis (S is a *continuous pivotal statistic*). Suppose we can generate N $i.i.d.$ replications, S_j, $j = 1 , \ldots , N$, of this test statistic under the null hypothesis. Compute

$$\widehat{G}_N(S_0) = \frac{1}{N}\sum_{j=1}^{N} I_{[0,\infty]}\left(S_j - S_0\right), \quad I_A(z) = \begin{cases} 1, & if\ z \in A \\ 0, & if\ z \notin A \end{cases} .$$

In other words, $N\widehat{G}_N(S_0)$ is the number of simulated statistics which greater or equal to S_0, and provided none of the simulated values S_j, $j = 1 , \ldots , N$, is equal to S_0,

$$\widehat{R}_N(S_0) = N - N\widehat{G}_N(S_0) + 1 \tag{3.1}$$

gives the rank of S_0 in the series $S_0, \ S_1, \ ..., \ S_N$.[11] Then the test's critical region corresponds to

$$\widehat{p}_N(S_0\ ) \leq \alpha\ , \ 0 < \alpha < 1. \tag{3.2}$$

where

$$\widehat{p}_N(x) = \frac{N\widehat{G}_N(x) + 1}{N + 1}\ . \tag{3.3}$$

The latter expression gives the *empirical probability* that a value as extreme or more extreme than S_0 is realized if the null is true. Hence $\widehat{p}_N(S_0\ )$ may be viewed as MC p-value.

Note that the MC decision rule may also be expressed in terms of $\widehat{R}_N(S_0)$. Indeed the critical region

$$\frac{N\widehat{G}_N(S_0\ ) + 1}{N + 1} \leq \alpha$$

is equivalent to

$$\widehat{R}_N(S_0) \geq (N + 1)\left(1 - \alpha\right) + 1\ . \tag{3.4}$$

[11] The subscript N in the notation adopted here may be misleading. We emphasize that $\widehat{R}_N(T_0)$ gives the rank of S_0 in the $N+1$ dimensional array $S_0, \ S_1, \ ..., \ S_N$. Throughout this section N refers to the number of MC replications.

In other words, for 99 replications a 5% MC test is significant if the rank S_0 in the series S_0 , S_1 , ..., S_N is at least 96, or informally, if S_0 lies in the series top 5% percentile.

> **We are now faced with the immediate question: does the MC test just defined achieve size control?**

If the null distribution of S is nuisance-parameter-free and $\alpha(N+1)$ is an integer, the critical region (3.2) is provably exact, in the sense that

$$\mathsf{P}_{(H_0)}\left[\widehat{p}_N(S_0\ ) \leq \alpha\right] = \alpha$$

or alternatively

$$\mathsf{P}_{(H_0)}\left[\widehat{R}_N(S_0) \geq (N+1)\,(1-\alpha)+1\right] = \alpha\,.$$

The proof is based on the following theorem concerning the distribution of the ranks associated with a finite dimensional array of exchangeable variables; see Dufour (1995) for a more formal statement of the theorem and related references.

3.1.1 Theorem *Consider an $M \times 1$ vector of exchangeable real random variables $(Y_1,\ ...,\ Y_M)$ such that $\mathsf{P}[Y_i = Y_j] = 0$ for $i \neq j$, and let R_j denote the rank of Y_j in the series $Y_1,\ ...,\ Y_M$. Then*

$$\mathsf{P}\left[\tfrac{R_j}{M} \geq z\right] = \tfrac{I[(1-z)M]+1}{M}\,, \quad 0 < z \leq 1\,. \tag{3.5}$$

where $I(x)$ is the largest integer less than or equal to x.

If S is continuous pivotal statistic, it follows from the latter result that

$$\mathsf{P}_{(H_0)}\left[\widehat{R}_N(S_0) \geq (N+1)\,(1-\alpha)+1\right]\,.$$

Indeed, in this case, the observed test statistic and the simulated statistic are exchangeable if the null is true. **Here it is worth recalling that the S_j's must be simulated imposing the null.** Now using (3.5), it is easy to show that $\mathsf{P}_{(H_0)}\left[\widehat{R}_N(S_0) \geq (N+1)\,(1-\alpha)+1\right] = \alpha$, provided N is chosen so that $\alpha(N+1)$ is an integer.

We emphasize that the sample size and the number of replications are explicitly taken into consideration in the above arguments. No central limit theorems have been used so far to justify the procedure just described.

It will be useful at this stage to focus on a simple illustrative example. Consider the Jarque and Bera normality test statistic,[12]

$$JB = n\left[\frac{1}{6}(Sk)^2 + \frac{1}{24}\,(Ku-3)^2\right]\,,$$

[12] See Section 2.3.2 for a formal presentation of the model and test statistics. Some equations are redefined here for convenience.

in the context of the linear regression model

$$Y = X\beta + u\,.$$

The MC test based on JB and N replications may be obtained as follows.

- Calculate the constrained OLS estimates $\widehat{\beta}$, s and the associated residuals $\widehat{u}$.
- Obtain the Jarque-Bera statistic based on s and $\widehat{u}$ and denote it $JB^{(0)}$.
- Treateing s as fixed, repeat the following steps for $j = 1, \ldots, N$:
  - ▶ draw an $(n \times 1)$ vector $\widetilde{u}^{(j)}$ as $i.i.d.\ N(0, s^2)$;
  - ▶ obtain the simulated independent variable $\widetilde{Y}^{(j)} = X\widehat{\beta} + \widetilde{u}^{(j)}$;
  - ▶ regress $\widetilde{Y}^{(j)}$ on X ;
  - ▶ derive the Jarque-Bera statistic $\widetilde{JB}^{(j)}$ associated with the regression of $\widetilde{Y}^{(j)}$ on X.
- Obtain the rank $\widehat{R}_N(JB^{(0)})$ in the series $JB^{(0)}$, $JB^{(1)}$, ..., $JB^{(N)}$..
- Reject the null if $\widehat{R}_N\left(JB^{(0)}\right) \geq (N+1)(1-\alpha) + 1$.

Furthermore, a MC p-value may be obtained as $\widehat{p}_N(S_0\ ) = [N + 1 - \widehat{R}_N(S_0)]/(N+1)$.

3.2. Monte Carlo tests in the presence of nuisance parameters

In Dufour (1995), we discuss extensions of MC tests when nuisance parameters are present. We now briefly outline the underlying methodology. In this section, n refers to the sample size and N the number of MC replications.

Consider a test statistic S for a null hypothesis H_0, and suppose the null distribution of S depends on an unknown parameter vector θ.

- From the observed data, compute:

(i) the test statistic S_0, and
(ii) a restricted consistent estimator $\widehat{\theta}_n^0$ of θ .

- Using $\widehat{\theta}_n^0$, generate N simulated samples and, from them, N simulated values of the test statistic. Then compute $\widehat{p}_N(S_0|\widehat{\theta}_n^0)$, where $\widehat{p}_N(x|\overline{\theta})$ refers to $\widehat{p}_N(x)$ based on realizations of S generated given $\theta = \overline{\theta}$ and $\widehat{p}_N(x)$ is defined in (3.3).
- A MC test may be based on the critical region

$$\widehat{p}_N(S_0|\widehat{\theta}_n^0) \leq \alpha\,,\ \alpha \leq 0 \leq 1\,.$$

▶ For further reference, we denote the latter procedure a **local Monte Carlo (LMC) test**.

Under general conditions, this LMC test has the correct level asymptotically (as $n \to \infty$), *i.e.,* under H_0,

$$\lim_{n\to\infty} \left\{ \mathsf{P}[\widehat{p}_N(S_0|\widehat{\theta}^0_n) \leq \alpha] - \mathsf{P}[\widehat{p}_N(S_0|\theta) \leq \alpha] \right\} = 0 \; . \tag{3.6}$$

In particular, these conditions are usually met whenever the test criterion involved is asymptotically pivotal. We emphasize that no asymptotics on the number of replication is required to obtain (3.6).

- To obtain an exact critical region, the MC p-value ought to be maximized with respect to the intervening parameters. Specifically, in Dufour (1995), we show that the test [henceforth called a **maximized Monte Carlo (MMC) test**] based on the critical region

  $$\sup_{\theta\, \in\, M_0} [\widehat{p}_N(S_0|\theta)] \leq \alpha \tag{3.7}$$

  where M_0 is the subset of the parameter space compatible with the null hypothesis (*i.e.*, the nuisance parameter space) is exact at level α.

- The LMC test procedure is closely related to a parametric bootstrap, with however a fundamental difference. Whereas bootstrap tests are valid as $N \to \infty$, the number of simulated samples used in MC tests is explicitly taken into account.

- Further the LMC p-value may be viewed as exact in a *liberal* sense, *i.e.* if the LMC fails to reject, we can be sure that the exact test involving the maximum p-value is not significant at level α.

In practical applications of exact MMC tests, a global optimization procedure is needed to obtain the maximal randomized p-value in (3.7). We use the simulated annealing (SA) algorithm [Corana, Marchesi, Martini, and Ridella (1987), Goffe, Ferrier, and Rogers (1994)].

To conclude this section, we consider another application of MC tests which is useful in the context of boundedly pivotal statistics. Using the above notation, the statistic at hand S is boundedly pivotal if it is possible to find another statistic S^* such that

$$S \leq S^*, \quad \forall \theta \in \Theta_0 \, , \tag{3.8}$$

and S^* is pivotal under the null. Let c and c^* refer to the α size-correct cut-off points associated with S and S^*. As emphasized earlier, inequality (3.8) implies that c^* may be used to define a critical region for S. The resulting test will be the correct level and may be viewed as **conservative** in the following sense: if the test based on c^* is significant, we can be sure that the exact test involving the (unknown!) c is significant at level α. The main point here is that it is easier to calculate c^*, because S^* is pivotal, whereas S is nuisance-parameter dependant. Of course, this presumes that the null exact distribution of S^* is known and tractable; see Dufour (1989, 1990) for the underlying theory and several illustrative examples. Here we argue that the MC test technique may be used to produce simulation-based conservative p-values based on S^* even if the analytic null distribution of S^* is unknown or complicated (but may be simulated). The procedure involved is the same as above, except that the S^* rather than S is evaluated from the simulated samples, as follows.

Let S_0 denote the observed test statistic S. By Monte Carlo methods and for a given number N of replications, generate S_j^*, $j = 1 \, , \, \ldots \, , \, N$, independent realizations of the bounding statistic S^*, under the null hypothesis. Obtain

$$\widehat{G}_N^*(S_0) = \frac{1}{N} \sum_{j=1}^{N} I_{[0,\infty]} \left(S_j^* - S_0\right)$$

and

$$\widehat{R}_N^*(S_0) = N - N\widehat{G}_N(S_0) + 1$$

where $\widehat{R}_N^*(S_0)$ gives the rank of S_0 in the series $S_0, S_1^*, \ldots, \; S_N^*$ .Then the LMC critical region corresponds to

$$\widehat{R}_N^*(S_0) \geq (N+1)\,(1-\alpha) + 1.$$

We denote the latter procedure a Bound MC (BMC) test.

A sound test strategy would be to perform the bounds tests first and, on failure to reject, to apply randomized tests. We recommend the following computationally attractive exact α test procedure:

1. compute the test statistic from the data;

2. if a bounding criterion is available, compute a BMC p-value; reject the null if: BMC p-value $\leq \alpha\,$;

3. if the observed value of the test statistic falls in the BMC acceptance region, obtain a LMC p-value; declare the test not significant if: LMC p-value $> \alpha\,$;

4. if the LMC p-value $\leq \alpha <$ BMC p-value, obtain the MMC p-value and reject the null if the latter is less than or equal to α.

4. Monte Carlo tests: econometric applications

4.1. Pivotal statistics

In Dufour and Kiviet (1996, 1998), Kiviet and Dufour (1997), Dufour, Farhat, Gardiol, and Khalaf (1998), Dufour and Khalaf (1998a, 1998c), Bernard, Dufour, Khalaf, and Genest (1998), Saphores, Khalaf, and Pelletier (1998), several applications of MC tests based on pivotal statistics are presented. The problems considered include the following ones.

- Normality tests [Dufour, Farhat, Gardiol, and Khalaf (1998)]
  - ▶ Kolmogorov-Smirnov's test
  - ▶ the Anderson-Darling test
  - ▶ Cramer-von Mises' test

  - the Shapiro-Wilk test
  - the Shapiro-Francia test
  - Weisberg-Binham's test
  - D'Agostino's test
  - the Filliben test
  - The Jarque-Bera LM test
  - the skewness and kurtosis coefficients

- Heteroskedasticity tests [Bernard, Dufour, Khalaf, and Genest (1998)]
  - Glejser's Wald-type tests;
  - Ramsey's versions of the Bartlett test;
  - Breusch-Pagan-Godfrey Lagrange multiplier (LM) test;
  - White's general test;
  - Koenker's studentized test
  - Szroeter's class of tests;
    - In this class of tests, we find that a Goldfeld-Quandt type test proposed by Szroeter performs particularly well in terms of power. The test takes the form of a Goldfeld-Quandt statistic where the variance estimates are obtained by partitioning the OLS residual vector from a single regression on the whole sample. Clearly, the criterion is not F distributed because the variance estimates so obtained are not independent. Yet it can be shown that it is pivotal, hence an exact MC version of this test is easily obtained. The fact that one single regression is needed may explain the superior power properties of this test.
  - Harrison and McCabe's test
  - Engle's LM test for ARCH and GARCH
  - Lee and King's LM-type test against ARCH and GARCH [Lee and King (1993)]
    - This test takes into account the one sided nature of the alternative hypothesis at hand (non-negativity of variance). We show that the MC version of this test outperforms the MC Engle test; the improvement in power is important.
  - extensions of the Cochrane and Hartley tests
    - The Hartley-type test takes the form $\max(s_i^2)/\min(s_i^2)$ where s_i^2 are estimates of the variance from relevant sub-samples (suggested by the alternative hypotheses). The statistic is intuitively appealing, yet the available tables of critical points are derived for the no-covariates case. We show that the MC version of this test may outperform the LR test. The example considered focuses on grouped heteroskedasticity, where the LR test requires an iterative maximization procedure and is clearly more expensive, particularly form the point of vue of simulation-based tests.

- ▶ Tests for break in variance at unknown points
    - - The tests proposed are of the sup(LM) and sup(LR) type. No trimming is required, and exact MC critical points are obtained. We show that as expected, the size of these tests is correct, regardless of the break date.

- Tests based on autocorrelations [Saphores, Khalaf, and Pelletier (1998)]
  - ▶ the Ljung-Box test
  - ▶ the variance ratio test.

In connection, it is worth mentioning that the MC test procedure applied to the Durbin-Watson test for AR(1) disturbances solves the inconclusive region problem.

The reader will find in the above papers simulation results which show clearly that the technique of Monte Carlo tests completely corrects often important size distortions due to poor large sample approximations. Now to illustrate the feasibility of MMC tests and the usefulness of BMC tests, we will focus on two examples involving nuisance parameters.

4.2. Monte Carlo tests in the presence of nuisance parameters: examples from the multivariate regression model

In this section, we review important distributional results from Dufour and Khalaf (1998a, 1998b) pertaining to LR test criteria in the MLR (reduced form) model. The model was introduced in Section 2.3.4. For convenience, we rewrite the system in stacked form:

$$y = (I_p \otimes X)\pi + v \tag{4.1}$$

where $y = vec(Y),\ \pi = vec(B),\ v = vec(U)$. Further, we suppose the disturbances have a know distribution up to a non-singular matrix. The distributional assumptions are formally set in Dufour and Khalaf (1998a, 1998c). Consider general restrictions on q^* independent linear transformations of π, of the form

$$H_{01} : R\pi \in \Delta_0 \tag{4.2}$$

where R is a $q^* \times kp$ matrix and Δ_0 is a non-empty subset of $\mathbb{R}^{q^*}$. This characterization of the hypothesis includes linear restrictions, both within and across equations, and allows for nonlinear as well as inequality constraints.

The LR criterion to test H_{01} is $n\ln(\Lambda)$, where

$$\Lambda = \frac{|\hat{\Sigma}_{01}|}{|\hat{\Sigma}|} \tag{4.3}$$

with $\hat{\Sigma}_{01}$ and $\hat{\Sigma}$ being the restricted and unrestricted ML estimators of Σ in (4.1). In the statistics literature, Λ^{-1} is often called the Wilks criterion. Dufour and Khalaf (1998a) show that the null distribution of Λ depends on nuisance parameters, yet it is boundedly pivotal. To see the point, consider restrictions of the form

$$H_{02} : Q\Pi C = D \tag{4.4}$$

such that $H_{02} \subseteq H_{01}$, where Q is a $q \times k$ matrix of rank q and C is a $p \times c$ matrix of rank c. Linear restrictions that decompose into the latter specific form are called *uniform linear* (UL) in the MLR literature. Let $\Lambda^c(q, c)$ be the reciprocal of the Wilks criterion for testing the latter restrictions. Then the distribution of Λ is bounded by the distribution of $\Lambda^c(q, c)$. Specifically, in Dufour and Khalaf (1998a), it is shown that:

(i) the null distribution of the LR statistic for uniform linear hypothesis involves no nuisance parameters and may easily be obtained by simulation;

(ii) under the null, $\mathsf{P}[\Lambda \geq \lambda^c_\alpha(q, c)] \leq \alpha$ for all $0 < \alpha < 1$, where $\lambda^c_\alpha(q, c)$ is determined such that $\mathsf{P}[\Lambda^c(q, c) \geq \lambda^c_\alpha(q, c)] = \alpha$.

The underlying distributional conditions are appreciably less restrictive than those of traditional multivariate analysis of variance which require normal errors.

4.2.1. Hypothesis testing in SURE models [13]

The results for the MLR model provide interesting applications for systems inference in SURE model. Let us now consider the following p equation SURE model:

$$Y_i = X_i\beta_i + U_i\,, \;\; i = 1,\, \ldots\,,\, p\,, \tag{4.5}$$

where X_i is a $n \times k_i$ full-column rank matrix of fixed regressors and $U_1, U_2\,,\,\ldots\,, U_p$ satisfy the distributional assumptions set above. Let

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ Y_p \end{bmatrix}, \; X^* = \begin{bmatrix} X_1 & 0 & \cdot & 0 \\ 0 & X_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & X_p \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \beta_p \end{bmatrix}, \; u = \begin{bmatrix} U_1 \\ U_2 \\ \cdot \\ U_p \end{bmatrix}.$$

Then an alternative compact representation of the model is

$$y = X^*\beta + u\,. \tag{4.6}$$

In this case, β is $k^* \times 1$ with $k^* = \sum_{i=1}^{p} k_i$. Note that the SURE model corresponds to a MLR system imposing exclusion restrictions; for further reference, we call this model the "nesting" MLR model. See Dufour and Khalaf (1998a) for an explicit formulation of the relation between both models. We consider the problem of testing a general hypothesis of the form

$$H_0 : C^*\beta \in \Delta^*_0 \tag{4.7}$$

[13] This section is based on Dufour and Khalaf (1998a).

where C^* is a full row-rank $v_0^* \times k^*$ matrix, Δ_0^* is a non-empty subset of $\mathbb{R}^{v_0^*}$. We first restate H_0 in terms of the MLR model which includes (4.6) as a special case, so that it incorporates the SURE exclusion restrictions. The associated LR statistic is

$$LR = n\ln(\Lambda)\ , \quad \Lambda = |\widehat{\Sigma}_0|/|\widehat{\Sigma}| \tag{4.8}$$

where $\widehat{\Sigma}_0$ and $\widehat{\Sigma}$ are the restricted and unrestricted SURE MLE. For the purpose of deriving the conservative bound, we also consider

$$LR^* = n\ln(\Lambda^*), \quad \Lambda^* = |\widehat{\Sigma}_0|/|\widehat{\Sigma}_u| \tag{4.9}$$

where $\widehat{\Sigma}_u$ is the unconstrained estimate of Σ in the "nesting" MLR model. As it stands, the testing problem may be approached as presented above. The following example illustrates the construction of the bounding statistic. Consider the three equations system

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11}X_1 + U_1\ , \\ Y_2 &= \beta_{20} + \beta_{22}X_2 + U_2\ , \\ Y_3 &= \beta_{30} + \beta_{33}X_3 + U_3\ , \end{aligned} \tag{4.10}$$

and the hypothesis

$$H_0 : \beta_{11} = \beta_{22} = \beta_{33}\ .$$

In the framework of the corresponding MLR model,

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + U_1\ , \\ Y_2 &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + U_2\ , \\ Y_3 &= \beta_{30} + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_3 + U_3\ , \end{aligned} \tag{4.11}$$

H_0 is equivalent to the joint hypothesis

$$H_0^* : \beta_{11} = \beta_{22} = \beta_{33} \text{ and } \beta_{12} = \beta_{13} = \beta_{21} = \beta_{23} = \beta_{31} = \beta_{32} = 0\,.$$

In order to use the above results on the conservative bound, we need to construct a set of UL restrictions that satisfy the hypothesis. It is easy to see that the constraints setting the coefficients $\beta_{ij},\ i,\ j\ =\ 1\ ,\ldots,\ 3$, to specific values meet this purpose. All that remains is to calculate associated Wilks' statistics conforming with these restrictions. Note that the statistic just derived serves to bound both LR and LR^*.

The algorithm for performing MC tests based on LR^*, at 5% level with 99 replications, can be described as follows.

- Compute $\widehat{\Sigma}_0$ and $\widehat{\Sigma}$, the restricted and unrestricted SURE (iterative) MLE .
- Compute $\widehat{\Sigma}_u$ is the unconstrained (OLS) estimate of Σ in the "nesting" MLR model.
- Compute $\Lambda^* = |\widehat{\Sigma}_0|/|\widehat{\Sigma}_u|$ and $LR^* = n\ln(\Lambda^*)$.

- Draw 99 realizations from a multivariate $(n, 3, I)$ normal distribution: $U^{(1)}, U^{(2)}$, ... , $U^{(p)}$ and store.

- Consider the linear constraints

$$H_{02}: \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{10} & \beta_{20} & \beta_{30} \\ \beta_{11} & \beta_{21} & \beta_{31} \\ \beta_{12} & \beta_{22} & \beta_{32} \\ \beta_{13} & \beta_{23} & \beta_{33} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_{11} & 0 & 0 \\ 0 & \widehat{\beta}_{22} & 0 \\ 0 & 0 & \widehat{\beta}_{33} \end{bmatrix}$$

where $\widehat{\beta}_{11} = \widehat{\beta}_{22} = \widehat{\beta}_{33}$ are the constrained SURE estimates calculated in the first step.

- Call the bounde MC procedure $BMC(\theta)$, described below, for $\theta = vec(chol(\widehat{\Sigma}_0))$. The Cholesky decomposition is used as an argument to impose positive definiteness. The output is the BMC p-value. Reject the null if the latter is $\leq$ .05 and STOP.

- Otherwise, call the procedure $MC(\theta)$, also described below, for $\theta = vec(chol(\widehat{\Sigma}_0))$. The output is the LMC p-value. Declare the test not significant if the latter $exceeds$ .05 and STOP.

- Otherwise, call the maximization algorithm for the function $MC(\theta)$ using $\theta = vec(chol(\widehat{\Sigma}_0))$ as a starting value. Obtain the MMC p-value and reject the null if the latter is $\leq$ .05. Note: if only a decision is required, the maximization algorithm may be instructed to exit once as soon as a value larger than .05 is attained. This may save considerable computation time.

▶ Description of the procedure $BMC(\theta)$:

- Construct a triangular Ω from θ (this gives the Cholesky decomposition of the variance which will be used to generate the simulated model).

- Do for $j = 1, \ldots, N$ (independently)

  - ⧫ Generate the random vectors $Y_1^{(j)}$ $Y_2^{(j)}$ $Y_3^{(j)}$ conformably with the nesting MLR model, using the restricted SURE coefficient estimates, the observed regressors, and Ω.

  - ⧫ Estimate the MLR model with the observed regressors as dependant variable, and $Y_1^{(j)}$ $Y_2^{(j)}$ $Y_3^{(j)}$ as independent variables: obtain the unrestricted estimates and the estimates imposing H_{02}.

  - ⧫ From these estimates, form the bounding statistics $LR_c^{(j)}$ and store.

- Obtain the rank of LR^* in the series LR^*, $LR_c^{(1)}$, ..., $LR_c^{(99)}$.

- This yields a BMC p-value as described above which is the output of the procedure.

▶ $\boxed{\text{Description of the procedure } MC(\theta)\text{:}}$

- Construct a triangular Ω from θ (this gives the Cholesky decomposition of the variance which will be used to generate the simulated model).
- Do for $j = 1, \ldots, N$ (independently)
    - ⧫ Generate the random vectors $Y_1^{(j)}\ Y_2^{(j)}\ Y_3^{(j)}$ conformably with the nesting MLR model, using the restricted SURE coefficient estimates, the observed regressors, and Ω.
    - ⧫ Estimate the MLR model with the observed regressors as dependant variable, and $Y_1^{(j)}\ Y_2^{(j)}\ Y_3^{(j)}$ as independent variables: obtain the unrestricted estimates and the estimates imposing H_0.
    - ⧫ From these estimates, form the statistics $LR^{*(j)}$ and store.
- Obtain the rank of LR^* in the series LR^*, $LR^{*(1)}$, ..., $LR^{*(99)}$.
- This yields a MC p-value as described above which is the output of the procedure.

▶ Description of the maximization algorithm (here: SA).

SA starts from an initial point, say $\theta = vec(chol(\widehat{\Sigma}_0))$, and sweeps the parameter space (user defined) at random. An *uphill* step is always accepted while a downhill step may be accepted; the decision is made using the Metropolis criterion. The direction of all moves is determined by probabilistic criteria. As it progresses, SA constantly adjusts the step length so that *downhill* moves are less and less likely to be accepted. In this manner, the algorithm escapes local optima and gradually converges towards the most probable area for optimizing. SA is robust with respect to non-quadratic and even non-continuous surfaces and typically escapes local optima. The procedure is known not to depend on starting values. Most importantly, SA readily handles problems involving a fairly large number of parameters.

In Dufour and Khalaf (1998a), we report the results of a simulation experiment designed according to this example. In particular, we examine the performance of LMC and BMC tests. We also return to the simulation experiment dealing with of UL in MLR contexts (see Section 2.3.4) and show that the Bartlett correction may fail, the MC test procedure achieves perfect size control.

4.2.2. Hypothesis tests in the simultaneous equation model [14]

This section discusses tests on structural parameters in SE models. We consider here the problem of testing linear restrictions in a LI framework; the notation for this case is set in Section 2.3.1 and is reproduced here for convenience:

$$\begin{aligned} y &= Y\beta + X_1\gamma_1 + u = Z\delta + u\,, \\ Y &= X_1\Pi_1 + X_2\Pi_2 + V\,, \end{aligned} \tag{4.12}$$

[14] This section is based on Dufour and Khalaf (1998b)

where Y and X_1 are $n \times m$ and $n \times k$ and X_2 refers to the excluded exogenous variables. The associated LI reduced form is

$$\begin{aligned} \left[\begin{array}{cc} y & Y \end{array} \right] &= X\Pi + \left[\begin{array}{cc} v & V \end{array} \right] \;\; \Pi = \left[\begin{array}{cc} \pi_1 & \Pi_1 \\ \pi_2 & \Pi_2 \end{array} \right], & (4.13) \\ \pi_1 &= \Pi_1\beta + \gamma_1, \;\; \pi_2 = \Pi_2\beta. & (4.14) \end{aligned}$$

We shall restrict attention to hypotheses that set several structural coefficients to specific values. More precisely, we consider in turn hypotheses of the form:

$$H_0 : \beta_i = \beta_i^0 \,, \tag{4.15}$$

$$H_0' : \beta_{1i} = \beta_{1i}^0 \,, \tag{4.16}$$

where $\beta_i = (\beta_{1i}', \beta_{2i}')'$ and β_{1i} is $m_{1i} \times 1$.

When the model is identified, (4.15) corresponds to the following restrictions

$$\Pi_{2i}\beta_i^0 = \pi_{2i} \tag{4.17}$$

or equivalently,

$$S_1 \left[\begin{array}{cc} \pi_{1i} & \Pi_{1i} \\ \pi_{2i} & \Pi_{2i} \end{array} \right] \left[\begin{array}{c} 1 \\ -\beta_i^0 \end{array} \right] = 0 \tag{4.18}$$

where

$$S_1 = \left[O_{(k-k_i,\, k_i)},\, I_{(k-k_i)} \right]$$

and $O_{(s,j)}$ denotes a zero $s \times j$ matrix. Let $\hat{\Sigma}_0$ and $\hat{\Sigma}$ be the error covariance LIML estimates imposing and ignoring (4.17), where the latter corresponds to the unrestricted reduced form. Further, let $\hat{\Sigma}_{LI}$ denote the LIML error covariance estimate imposing the exclusion restrictions implied by the structure. Conformably with the notation set above, define

$$\Lambda_{LI} = \frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_{LI}|} \,, \tag{4.19}$$

$$\Lambda^c(k - k_i, 1) = \frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}|} . \tag{4.20}$$

Following the arguments of Section 3.2, we see that the distribution of Λ_{LI} is bounded by the distribution of $\Lambda^c(k - k_i, 1)$.

The limited information LR statistic (LR_{LI}) may be obtained as $n \ln(\Lambda_{LI})$. Whereas $n[\ln(\Lambda_{LI})]$ has a $\chi^2(m_i)$ asymptotic distribution only under identification assumptions, $n[\ln(\Lambda^c(k - k_i, 1))]$ is asymptotically distributed as $\chi^2(k - k_i)$ whether the rank condition holds or not. The asymptotic distribution of the LR_{LI} statistic is thus bounded by a $\chi^2(k - k_i)$ distribution independently of the conditions for identification. This result was derived under *local-to-zero* asymptotics in Nelson, Startz, and Zivot (1996) and Wang and Zivot (1996) for the special case where β_i is scalar. Furthermore, exact bounds based on the $F(k - k_i, n - k)$ distribution may also be derived for this problem if normality is imposed. Indeed, as pointed out in Stewart (1997),

$$\frac{q[\Lambda^c(q,1)-1]}{n-k} \sim F(q, n-k) \tag{4.21}$$

where $\Lambda^c(q,1)$ is the reciprocal of the Wilks statistics for testing uniform linear restrictions of the form $Q\Pi C = D,\ rank(Q) = q,\ rank(C) = c = 1$.

The important thing to note regarding the latter bound is that it relates to the well known Anderson-Rubin (AR) statistic. It is straightforward to show using the results on UL hypotheses in Dufour and Khalaf (1998a) and Stewart (1997) that the AR statistic associated with $H_0 : \beta_i = \beta_i^0$ corresponds to a monotonic transformation of the LR criterion for testing the UL hypothesis $\Pi_{2i}\beta_i^0 = \pi_{2i}$ against an unrestricted alternative.

Let us now consider the hypothesis (4.16). On partitioning $\Pi_{1i} = [\Pi_{11i}, \Pi_{12i}]$ and $\Pi_{2i} = [\Pi_{21i}, \Pi_{22i}]$ conformably with $\beta_i = (\beta_{1i}', \beta_{2i}')'$ the corresponding reduced form restrictions may be expressed as

$$S_2 \begin{bmatrix} \pi_{1i} & \Pi_{11i} & \Pi_{12i} \\ \pi_{2i} & \Pi_{21i} & \Pi_{22i} \end{bmatrix} \begin{bmatrix} 1 \\ -\beta_{1i}^0 \\ -\beta_{2i} \end{bmatrix} = 0 \tag{4.22}$$

where

$$S_2 = \left[O_{(k-k_i\ ,\ k_i)}\,, I_{(k-k_i)}\right] \ .$$

Let Λ_{LI} be the reciprocal of the Wilks statistic for testing (4.22) against the restrictions implied by the structure. The nonlinearities in connection with (4.22) stem from the fact that β_{2i} is unknown. However, the special case of (4.22) that corresponds to specific (unknown) values of β_{2i} takes the UL form. Let $\Lambda_\alpha^c(k-k_i,\ 1)$ denote the reciprocal of the Wilks statistic for testing these UL restrictions against an unrestricted alternative. Then conservative bounds for Λ_{LI} can be obtained from the statistic $\Lambda^c(k-k_i, 1)$ or the $F(k-k_i, n-k)$ when applicable. In Dufour and Khalaf (1998b), we present simulations which illustrate the performance of MC tests in this context. The parameters for the simulation experiments were presented in Section 3.1.

We have attempted to apply the MC test procedure to the Wald test for both hypotheses considered. In this case, the performance of the standard bootstrap was disappointing. The LMC Wald tests failed completely in near-unidentified conditions. Furthermore, in all cases examined, the Wald tests maximal randomized p-values were always one. This is a case where the MC procedure does not (and cannot) correct the performance of the test.

> *In other words, Wald statistics do not constitute valid pivotal functions in such models and it is even impossible to bound their distributions over the parameter space (except by the trivial bound* 0 *and* 1*).* [Dufour (1997)]

We conclude this section with a specific problems where the MC test strategy conveniently solves a difficult and non-standard distributional problems: combining non-independent test criteria.

4.3. Monte Carlo tests in non-standard test problems

4.3.1. Combining non-independent tests [15]

In the context of the SURE model (4.5), consider the hypothesis that Σ is diagonal:

$$H_0 : \Sigma = D_p(\sigma_i^2), \text{ for some vector } (\sigma_1 \, , \, ... \, , \, \sigma_p)' \, , \tag{4.23}$$

where $D_N(d_i)$ represent a diagonal matrix of dimension N, with $(d_1 \, , \, ... \, , \, d_N)$ along the diagonal. Several criteria are available for testing H_0 of which the most well known are the LR and the LM tests [Breusch and Pagan (1980)]. The asymptotic null distribution of both criteria is $\chi^2(p(p-1)/2)$. Dufour and Khalaf (1998c) show that LR and LM statistics are pivotal under the null, which implies that exact critical values can be obtained easily by MC techniques no matter how large the system is.

On the other hand, a finite sample exact independence test was developed by Harvey and Phillips (1980, HP); their procedure is applicable where the null hypothesis has the form

$$H_{01} : \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \Sigma_{11} \end{bmatrix} . \tag{4.24}$$

Specifically, they propose to use the usual F statistic for testing whether the coefficients on $\hat{V}_1$ are zero in the regression of y_1 on X_1 and $\hat{V}_1$, where $\hat{V}_1$ is the matrix of OLS residuals associated with the other equations in the system.

HP tests may be applied in the context of general diagonality tests; for example, one may assess in turn whether the disturbances in each equation are independent of the disturbances in all other equations. Such a sequence of tests however raises the problem of taking into account the dependence between multiple tests, a problem not solved by Harvey and Phillips (1982); in fact, the problem of testing (4.23) is not addressed at all in Harvey and Phillips (1982). Here we introduce several induced tests of (4.23) based on a set of simultaneous HP-type tests and suggest a simulation-based solution to the associated combination problem. The critical regions of conventional induced tests are usually computed using probability inequalities (*e.g.,* the well know Boole-Bonferroni inequality) which yields conservative p-values whenever non independent tests are combined [see, for example, Savin (1984), Folks (1984) and Dufour and Torrès (1998, 1996)]. Here, we propose to construct the induced tests such that size-corrected p-values can be readily obtained by simulation.

We examine two types of induced tests. The first one involves combining the p HP independence tests between the errors of each equation and those of the other equations in the system. The separate tests involved and the associated p-values are denoted $F(i, \, \bar{\imath})$ and $p(i, \, \bar{\imath}), \; i = 1, \, \ldots \, , \, p$, where

$$p(i, \, \bar{\imath}) = \mathsf{P}[F(i, \, \bar{\imath}) \geq F(p-1, \, n - k_i - p + 1] \, .$$

Alternatively, we propose to examine, in turn, whether u_1 is independent of $(u_2 \, , \, ... \, , u_p)$, or u_2 is independent of $(u_3 \, , \, ... \, , \, u_p)$ and so on so forth; a maximum of $(p-1)$ tests would be involved. In this case, we denote the sequential individual tests and corresponding p-values $F(i, \; i : p)$ and

[15] This section is based on the results in Dufour and Khalaf (1998c).

$p(i,\ i:p),\ i=1,\ \ldots\ ,\ p$, where

$$p(i,\ i:p) = \mathsf{P}[F(i,\ i:p) \geq F(p-1,\ n-k_i-p+i)].$$

For further reference, the former procedure is called a *joint* test, and the latter a *sequential* test. To obtain an α-level procedure, the most common practice in such situations involves Bonferroni-based tests which reject the null hypothesis when at least one of the separate tests is significant at level α_i, where the sum of the individual significance levels α_i overall the tests performed equals α. For the Bonferroni joint test, we shall use $\alpha_i = \alpha/p$. For the Bonferroni sequential test, we set $\alpha_1 = \alpha/2$, $\alpha_2 = \alpha/(2^2), \ldots, \alpha_{p-2} = \alpha/(2^{p-2})$, $\alpha_{p-1} = \alpha/(2^{p-2})$. We also discuss how one can perform multiple tests of size α based on the statistics $F(i,\ \bar{i})$ and $F(i,\ i:p)$. To do this, we propose the following combined test criteria:

$$F_{\text{JOINT, MIN}} = 1 - \min_p \left\{p(i,\ \bar{i})\right\}, \tag{4.25}$$

$$F_{\text{JOINT, PROD}} = 1 - \prod_{i=1}^{p} \left\{p(i,\ \bar{i})\right\}, \tag{4.26}$$

$$F_{\text{SEQ, MIN}} = 1 - \min_p \{p(i,\ i:p)\}, \tag{4.27}$$

$$F_{\text{SEQ, PROD}} = 1 - \prod_{i=1}^{p} \{p(i,\ i:p)\}. \tag{4.28}$$

The combined test involving the smallest p-value was suggested by Tippett (1931) and Wilkinson (1951). In the case of the joint test, an equivalent procedure may be obtained using the largest $F(i,\ \bar{i})$. The combination method using the product of the p-values goes back to Fisher (1932) and Pearson (1933). Both multiple test procedures as originally suggested are valid when the combined tests are independent. Here, we propose to apply the MC technique based on these criteria may be used to obtain induced tests of size α.

In order to assess the performance of the various procedures discussed above, a set of Monte Carlo experiments were conducted for a five equation model ($p = 5$) with $n = 25$. To assess the tests sizes, we also consider $n = 50, 100$. The design matrices X_i , $i = 1,\ \ldots\ ,\ p$, included a constant term, a set of k_m explanatory variables common to all equations and a set of regressors which differ between the different equations, where

$$k_m = \begin{cases} k_i/2, & \text{if } k_i \text{ is even} \\ (k_i+1)/2, & \text{if } k_i \text{ is odd} \end{cases}.$$

Hence, $k_m - 5(k_i - k_m - 1)$ distinct variables were used in each design and were generated using a multivariate normal distribution. The set of regressors were kept constant for all replications. The disturbances were generated from multivariate normal distributions. Several choices for the error covariance were considered. The matrix labelled Σ_1 as well as the regression coefficients are from the empirical example discussed in the paper. The other matrices were obtained by dividing certain elements of the Cholesky decomposition of Σ_1 by appropriate constants to decrease the covariance terms. Of course, the parameters under the null were obtained by setting the non-diagonal elements of Σ_1 to zero. The number of trials for the MC tests was set to 19 and 99 (N = 19, 99). The number of overall replications was 1000. The following table summarizes basic results.

Table 2. Empirical rejections of various independence tests

$n = 25$	Σ_0 (H_0)		Σ_1		Σ_2		Σ_3		Σ_4	
	ASY	MC	MC		MC		MC		MC	
Reps	-	19	19	99	19	99	19	99	19	99
LM	.105	.045	.998	1.0	.911	.954	.704	.794	.444	.500
LR	.267	.047	1.0	1.0	.961	.980	.746	.824	.428	.494
$F_{JOINT,MIN}$	-	.043	1.0	1.0	.925	.965	.632	.693	.360	.409
$F_{JOINT,PROD}$	-	.052	1.0	1.0	.944	.980	.714	.784	.382	.438
$F_{SEQ,MIN}$	-	.052	1.0	1.0	.963	.984*	.721	.799	.490	.562*
$F_{SEQ,MIN}$	-	.049	1.0	1.0	.846	.912	.562	.653	.368	.399
	Bonferroni Harvey-Phillips type tests									
F_{ALL}	.034		1.0		.963		.665		.356	
F_{SEQ}	.049		1.0		.896		.687		.316	

From the results in this table, we can make the following observations.

1. The asymptotic tests consistently overreject. In contrast, all MC tests achieve size control.

2. The size corrected tests perform quite well, although the LR-type tests seem rather superior.

3. The Bonferroni tests show relatively low power, as we would expect.

4. The MC induced tests based on the Harvey-Phillips statistics perform very well overall the parameter values considered. As expected, the Tippet/Wilkinson-type MC induced tests perform better than their Bonferroni counterparts. The power of the Fisher/Pearson MC induced test is generally higher than the Tippet/Wilkinson-type MC test. Furthermore, the former test displays good power properties with respect to the LR based MC tests and in some cases outperforms the MC LR tests.

4.3.2. Non-identified nuisance parameters[16]

The first example we discuss in this section is the problem of testing for the significance of jumps in the context of a jump-diffusion model. For econometric applications and references, see Saphores, Khalaf, and Pelletier (1998). Formally, consider the following model written, for convenience, in discrete time:

$$S_t - S_{t-1} = \mu + \sigma \xi_t + \sum_{i=1}^{n_t} \ln(Y_t) \, , \; t = 1, \, \ldots \, , \, T,$$

[16] This section draws on the results in Saphores, Khalaf, and Pelletier (1998) and Bernard, Dufour, Khalaf, and Genest (1998).

where $\xi \overset{iid}{\sim} N(0,1)$ and $\ln(Y) \overset{iid}{\sim} N(\theta, \delta^2)$ and n_t is the number of jumps which occur in the interval $\left[\ t-1,\ \ t\ \right]$; the arrival of jumps is assumed to follow a *Poisson* process with parameter λ. The associated likelihood function is as follows:

$$L_1 = -T\ln(\lambda) - \frac{T}{2}\ln(2\pi) + \sum_{t=1}^{T} \ln\left[\sum_{j=0}^{\infty} \frac{\lambda^j}{j!}\frac{1}{\sqrt{\sigma^2+\delta^2 j}}\exp\left(\frac{-(x_t-\mu-\theta j)^2}{2(\sigma^2+\delta^2 j)}\right)\right] .$$

The hypothesis of no jumps corresponds to $\lambda = 0$. It is clear that in this case, the parameters θ, δ^2 are not identified under the null, and hence, following the results of Davies (1977, 1987), the distribution of the associated LR statistic is non-standard and quite complicated. Although this problem is well recognized by now, a $\chi^2(3)$ asymptotic distribution is often (inappropriately) used in empirical applications of the latter LR test. See Diebold and Chen (1996) for related arguments dealing with structural change tests.

Let $\widehat{\mu}, \widehat{\sigma}^2$ denote the MLE under the null, i.e. imposing a Geometric Brownian Motion. Here we argue that in this case, the MC p-value calculated as described above, drawing $i.i.d.\ N(\widehat{\mu}, \widehat{\sigma}^2)$ disturbances (with $\widehat{\mu}$ and $\widehat{\sigma}^2$ taken as given) will not depend on θ and δ^2. This follows immediately from the implications of non-identification. Furthermore, the invariance to location and scale (μ and σ) is straightforward to see. Consequently, the MC test described in the context of pivotal statistics will yield exact p-values.

The problem of unidentified nuisance parameters is prevalent in econometrics. We next turn to another illustrative example: testing for ARCH-in-mean effects. Formally, consider the model:

$$\begin{aligned} Y_t &= X_t'\beta + h_t\phi + e_t\,, \\ e_{t|t-1} &\sim N(0\,,\,h_t^2)\,, \\ h_t^2 &= \alpha_0 + \alpha_1 e_{t-1}^2 + ... + \alpha_p e_{t-p}^2\,, \end{aligned}$$

where the predetermined variables x_t include a constant regressor and the notation $e_{t|t-1}$ denotes conditioning of information up to and including $t-1$. The hypothesis of no-ARCH is

$$H_0 : \alpha_1 = ... = \alpha_p = 0\,.$$

Let $\widehat{\beta}$ and $\widehat{\sigma}^2$ denote the OLS estimates from the regression of Y_t on X_t and $\widehat{e}_t$ the associated residuals. The LM test for a known ϕ is

$$LM(\phi) = \frac{1}{2+\phi^2}\gamma' W\left[W'W - \frac{\phi^2}{2+\phi^2}W'X(X'X)^{-1}X'W\right]^{-1} W'\gamma \qquad (4.29)$$

where γ is a $T \times 1$ vector with elements

$$\gamma_t = (-1) + \phi\widehat{e}_t/\widehat{\sigma}$$

and W is a $T \times (p+1)$ matrix with elements

$$W_t = (\ 1,\ \ \widehat{e}^2_{t-1},\ \ ...,\ \ \widehat{e}^2_{t-p}\ ).$$

In this case, it is also evident that under the null, the parameter ϕ is unidentified. Indeed, only (the intercept + ϕ) may be ”estimated” from the data. In practice, the latter estimate is substituted for ϕ to implement the LM test using a cut-off point from the $\chi^2(p)$. Bera and Ra (1995) discuss the application of the Davies sup-LM test to this problem and show that this leads to more reliable inference. It is clear however that the asymptotic distribution required is quite complicated. The MC test procedure may be applied to the sup-LM test. Simulation studies reported in Bernard, Dufour, Khalaf, and Genest (1998) show that this method works very well in terms of size and power.

4.3.3. Cointegration tests [17]

Consider the p-dimensional k-lags VAR

$$\Delta X_t = \Pi X_{t-1} + \sum_{j=1}^{k-1} \Gamma_j \Delta X_{t-j} + e_t$$

using standard notation, fixed initial values and i.i.d disturbances with covariance Σ. The hypothesis that there are r or fewer cointegrating vectors is

$$H_{(r)} : \text{rank}(\Pi) \leq r \,.$$

The associated test statistic is computed in terms of the square canonical correlations of $\widehat{\Delta X_t}$ and $\widehat{X}_{t-1}$ where $\widehat{\Delta X_t}$ and $\widehat{X}_{t-1}$ refer to residuals from the regression of ΔX_t and X_{t-1} on all lagged differences. Nielsen (1998) emphasizes that the problem of solving the underlying determinantal equation is invariant to non-singular linear transformations of the data. Correspondingly, the null distribution of the Johansen test statistic is scale-invariant.

As it is well know, the asymptotic null distributions underlying the specialized tables of critical points provide for use with the Johansen test are asymptotically similar. More specifically, the asymptotic distributions depend on the assumed number of unit roots. Nielsen argues that this result must be qualified in models with $k > 1$. Whereas the test of no-cointegration is "exact similar" in the case $k = 1$, the same test is not even asymptotically pivotal is higher order models.

The point is best illustrated with the model corresponding to $k = 2, p = 1$. In this case, Nielsen shows analytically and using simulations that the null distribution for the Johansen test depends importantly on the coefficient of the lagged variable Γ. To extract the analytic null distribution, Nielsen reparameterizes the model scaling the latter coefficient by the sample size. Furthermore, tables for the asymptotic distribution fixing $T\Gamma$ are provided and serve to illustrate the dependence on Γ. Whether these tables may be used for empirical applications is an open question. The author recognizes that in practice, Γ must be estimated from the data, which may cast doubt on the reliability of these tables. Related identification issues in the context of cointegration tests are also discussed in Dufour (1997).

In view of this, the MMC approach may be worth considering here. Note also that the results in Dufour and Khalaf (1998a) on the pivotal bound may provide useful applications to this test problem.

Finally, a word of caution about structural VAR models. Consider the two-dimensional structural VAR model

$$X_t = A(L)\varepsilon_t$$

and suppose the standard normalization and orthogonalization restrictions are imposed. It is also assumed that the second shock has no lung run effect on the first variable, *i.e.* $A_{12}(1) = 0$. We are interested in testing

$$H_0 : a_{12k} = 0$$

[17] This section draws on the results in Nielsen (1998).

i.e. whether the response of the first variable to the second shock at lag k is zero. Faust and Leeper (1997) demonstrate that in this case, any size-correct test has power less than or equal to its size. The proof is based on the following arguments.

Consider a model satisfying $a_{12k} \neq 0$, and a test statistic which has maximum power at level α for this model. Specifically, this means that the probability that the test rejects the model (which we denote β) is higher than the probability that the test rejects any other model under the alternative. It is always possible to transform the model as follows:

$$Z_t = B(L)e_t$$

where

$$\begin{aligned} B(L) &= A(L)B\ ,\ e_t = B^{-1}\varepsilon_t \\ e_t &= B^{-1}\varepsilon_t \end{aligned}$$

and B is chosen such that $b_{12k} = 0$. It si clear that the modified model satisfies the null, and the rejection probability β is not affected by the transformation, since the modified model and the original model are observationally equivalent. It is also possible to modify the model further so that the long run neutrality condition is satisfied, yet the rejection probability is only very slightly altered. If the test is level-correct, then $\beta \leq \alpha$, because the modified model satisfies the null.

The MC procedure cannot correct this problem. As emphasized above, the MC test technique solves the problem of size control but even if size is controlled in this case, the test and any related confidence interval are not useful.

4.4. Confidence intervals from MC tests [18]

In this section, we reconsider the probit model discussed in Section 2.3.3 and propose two alternative methods to obtain relevant confidence intervals for the parameter ratio β_i/β_j. Both methods require inverting relevant test statistics for the hypothesis

$$H_0(\delta_0) : \beta_i/\beta_j = \delta_0\ . \tag{4.30}$$

In other words, the $(1-\alpha)-$ level confidence set corresponds to the values of δ_0 for which the test statistic considered fails to reject $H_0(\delta_0)$ at level α.

The first method is based on an approach proposed long ago by Fieller (1940, 1954) and relies on inverting a *t*-type statistic

$$t(\delta_0) = \frac{\widehat{\beta}_i - \delta_0\widehat{\beta}_j}{\mathrm{SE}(\widehat{\beta}_i - \delta_0\widehat{\beta}_j)} \tag{4.31}$$

using any consistent parameter and variance/covariance estimates including simulated maximum likelihood (SML) criteria. The above procedure is justified on asymptotic grounds. It is easy to show that the derived CI are not necessarily bounded; thus they will not suffer from the fundamental

[18] This section draws on the results in Bolduc-Dufour-Khalaf (1998).

limitations documented in Dufour (1997) which preclude valid inference. Indeed, inverting (4.31) implies solving a 2nd degree polynomial where unbounded sets are possible solutions.

To explore the feasibility of Fieller-type CI, we have applied the t-test based procedure to a trinomial logit model of travel demand analyzed in Ben-Akiva and Lerman (1985, Chapter 7). We consider a CI for the ratio β_3/β_5 where β_3 is the coefficient for "round trip travel time" and β_5 is the coefficient for "round trip travel cost" [see Ben-Akiva and Lerman (1985, p. 158)]. The *delta*-method yields

$$\left[\ -0.0002089\,,\quad 0.0023483\ \right] \tag{4.32}$$

whereas the Fieller methods gives

$$\left[\ -\infty\,,\quad -0.0151209\ \right] \cup \left[\ 0.0003947\,,\quad +\infty\ \right]. \tag{4.33}$$

Although the Fieller-type region is unbounded and may appear too wide and uninformative using traditional arguments, notice that (4.33) does not include zero. This implies that the ratio is statistically significantly different from zero; in terms of the hypothesis $\beta_3/\beta_5 = 0$, the Fieller CI is indeed quite informative. This contrasts with the inference based on (4.32), although both regions are derived using the same asymptotically normal parameter and variance/covariance estimates. This example illustrates a situation where the *delta*-method and our proposed method may be in conflict. As demonstrated in Section 2.3.1, the former method is suspect, whereas the latter is more likely to yield confidence sets immune to the severe size problems documented in Dufour (1997).

Similar non-standard confidence estimation techniques have gained popularity in recent econometric practice, particularly in non-regular inference problems. Indeed, Staiger and Stock (1997), Dufour and Jasiak (1994) and Stock and Wright (1997) derive inversion-based confidence sets to solve the poor instrument problem in instrumental regressions. See also Abdelkhalek and Dufour (1998) for related techniques in general equilibrium models.

We next suggest an alternative confidence estimation procedure based on a MC test. In view of the results (cited above) documenting the good performance of bootstrap based tests in discrete choice contexts, one would expect the associated confidence sets to work well. MMC techniques provide a formal strategy for dealing with nuisance parameter so that exact confidence sets may be obtained, regardless of the complicated non-linear structure of the model at hand. SML-based test criteria may be used. The procedure may be implemented as follows.

First modify the above notation: let S denote any valid ("simulatable") statistic for testing

$$H_0 : \beta_i - \delta_0\beta_j = 0$$

and let $(\beta_i,\, \beta_j,\, \theta)$ refer to the relevant intervening parameters (θ is a vector of extra nuisance parameters on which the distribution of S under H_0 may depend). Define

$$\widehat{p}_N(\beta_i,\, \beta_j,\, \theta)$$

as the $\alpha-$level MC p-value obtained following the lines described above. Conformably, let

$$\widetilde{p}_N\,(\delta_0) = \sup_{\beta_j,\theta} \widehat{p}_N(\delta_0\beta_j,\, \beta_j,\, \theta)\,.$$

Then a valid confidence interval δ_0 is $(q_L,\, q_U)$,where

$$q_L = \inf_{\delta_0}\{\delta_0 : \widetilde{p}_N(\delta_0) > \alpha\}\,, \quad q_U = \sup_{\delta_0}\{\delta_0 : \widetilde{p}_N(\delta_0) > \alpha\}\,.$$

Indeed, it can be shown from the arguments in Dufour (1995) that

$$\mathsf{P}[\delta \in (q_L,\, q_U)] \geq 1 - \alpha\,.$$

Clearly, MC procedures are computationally intensive. However, they guarantee the desired coverage. Bolduc, Dufour, and Khalaf (1998) propose to construct feasible algorithms for practical use and compare the relative merits of both the Fieller and MC procedure. Although we have discussed the method in the context of probit models, both procedures are clearly applicable in various other settings.

5. Conclusion

In this paper, we have demonstrated that finite sample concerns may arise in several empirically pertinent test problems. But, in many cases of interest, the MC test technique produces valid inference procedures no matter how small your sample is.

We have also emphasized that the problem of constructing a good test - although simplified - cannot be solved **just** using simulations. Yet in most examples we have reviewed, MC test techniques emerge as indispensable tools.

Beyond the cases covered above, it is worthwhile noting that the MC test technique may be applied to many other problems of interest. These include, for example, models where the estimators themselves are also simulation-based, *e.g.*, estimators based on indirect inference or involving simulated maximum likelihood. Furthermore, the MC test technique is by no means restricted to nested hypotheses. It is therefore possible to compare non-nested models using MC LR-type tests; assessing the success of this strategy in practical problems is an interesting research avenue.

Your data is valuable, and the statistical analysis you perform is often policy oriented. Why tolerate questionable p-values and confidence intervals, when exact or improved approximations are available?

References

ABDELKHALEK, T., AND J.-M. DUFOUR (1998): "Statistical Inference for Computable General Equilibrium Models, with Application to a Model of the Moroccan Economy," *Review of Economics and Statistics*, LXXX, 520–534.

ANDERSON, T. W., AND H. RUBIN (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63.

ATTFIELD, C. L. F. (1995): "A Bartlett Adjustment to the Likelihood Ratio Test for a System of Equations," *Journal of Econometrics*, 66, 207–223.

BARNARD, G. A. (1963): "Comment on 'The Spectral Analysis of Point Processes' by M. S. Bartlett," *Journal of the Royal Statistical Society, Series B*, 25, 294.

BARNDORFF-NIELSEN, O. E., AND P. BLAESILD (1986): "A Note on the Calculation of Bartlett Adjustments," *Journal of the Royal Statistical Society, Series B*, 48, 353–358.

BARTLETT, M. S. (1937): "Properties of Sufficiency and Statistical Tests," *Proceedings of the Royal Society of London A*, 160, 268– 282.

——— (1948): "A Note on the Statistical Estimation of Supply and Demand Relations from Time Series," *Econometrica*, 16, 323–329.

BEN-AKIVA, M., AND S. R. LERMAN (1985): *Discrete Choice Analysis: Theory And Application to Travel Demand.* The MIT Press, Cambridge, MA.

BERA, A. K., AND S. RA (1995): "A Test for the Presence of Conditional Heteroscedasticity Within ARCH-M Framework," *Econometric Reviews*, 14, 473–485.

BERNARD, J.-T., J.-M. DUFOUR, L. KHALAF, AND I. GENEST (1998): "Simulation-Based Finite-Sample Tests for Heteroskedasticity and ARCH Effects," Discussion paper, Département d'économique, Université Laval, and CRDE, Université de Montréal.

BERNDT, E. R. (1991): *The Practice of Econometrics: Classic and Contemporary*. Addison-Wesley, Reading (MA).

BIRNBAUM, Z. W. (1974): "Computers and Unconventional Test-Statistics," in *Reliability and Biometry*, ed. by F. Proschan, and R. J. Serfling, pp. 441–458. SIAM, Philadelphia, PA.

BOLDUC, D., J.-M. DUFOUR, AND L. KHALAF (1998): "Confidence Sets for Parameter Ratios in Discrete Choice Models," Discussion paper, CRDE, Université de Montréal, and GREEN, Université Laval.

BOUND, J., D. A. JAEGER, AND R. M. BAKER (1995): "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

BREUSCH, T. S., AND A. R. PAGAN (1980): "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47, 239–254.

CORANA, A., M. MARCHESI, C. MARTINI, AND S. RIDELLA (1987): "Minimizing Multimodal Functions of Continuous Variables with the 'Simulated Annealing' Algorithm," *ACM Transactions on Mathematical Software*, 13, 262–280.

D'AGOSTINO, R. B., AND M. A. STEPHENS (eds.) (1986): *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

DAVIDSON, R., AND J. G. MACKINNON (1999a): "Bootstrap Testing in Non-linear Models," *International Economic Review*, forthcoming.

——— (1999b): "Bootstrap Tests: How Many Bootstraps," *Econometric Reviews*, forthcoming.

——— (1999c): "The Size Distortion of Bootstrap Tests," *Econometric Theory*, 15, 361–376.

DAVIES, R. B. (1977): "Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 64, 247–254.

DAVIES, R. B. (1987): "Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative," *Biometrika*, 74, 33–43.

DAVISON, A., AND D. HINKLEY (1997): *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge (UK).

DIEBOLD, F. X., AND C. CHEN (1996): "Testing Structural Stability with Endogenous Break Point: A Size Comparison of Analytic and Bootstrap Procedures," *Journal of Econometrics*, 70, 221–241.

DUFOUR, J.-M. (1989): "Nonlinear Hypotheses, Inequality Restrictions, and Non-Nested Hypotheses: Exact Simultaneous Tests in Linear Regressions," *Econometrica*, 57, 335–355.

——— (1990): "Exact Tests and Confidence Sets in Linear Regressions with Autocorrelated Errors," *Econometrica*, 58, 475–494.

——— (1995): "Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics," Discussion paper, C.R.D.E., Université de Montréal.

——— (1997): "Some Impossibility Theorems in Econometrics, with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1389.

DUFOUR, J.-M., A. FARHAT, L. GARDIOL, AND L. KHALAF (1998): "Simulation-Based Finite Sample Normality Tests in Linear Regressions," *The Econometrics Journal*, 1, 154–173.

DUFOUR, J.-M., AND J. JASIAK (1994): "Finite Sample Inference Methods for Simultaneous Equations and Models with Unobserved and Generated Regressors," Discussion paper, C.R.D.E., Université de Montréal, 38 pages.

DUFOUR, J.-M., AND L. KHALAF (1998a): “Monte Carlo Tests for Contemporaneous Correlation of Disturbances in Multiequation SURE Models,” Discussion paper, C.R.D.E., Université de Montréal.

——— (1998c): “Simulation Based Finite and Large Sample Inference Methods in Multivariate Regressions and Seemingly Unrelated Regressions,” Discussion paper, C.R.D.E., Université de Montréal, 36 pages.

——— (1998b): “Simulation-Based Finite and Large Sample Inference Methods in Simultaneous Equations,” Discussion paper, C.R.D.E., Université de Montréal.

DUFOUR, J.-M., AND J. F. KIVIET (1996): “Exact Tests for Structural Change in First-Order Dynamic Models,” *Journal of Econometrics*, 70, 39–68.

——— (1998): “Exact Inference Methods for First-Order Autoregressive Distributed Lag Models,” *Econometrica*, 66, 79–104.

DUFOUR, J.-M., AND O. TORRÈS (1996): “Markovian Processes, Two-Sided Autoregressions and Exact Inference for Stationary and Nonstationary Autoregressive Processes,” Discussion paper, C.R.D.E., Université de Montréal, 28 pages.

——— (1998): “Union-Intersection and Sample-Split Methods in Econometrics with Applications to SURE and MA Models,” in *Handbook of Applied Economic Statistics*, ed. by D. E. A. Giles, and A. Ullah, pp. 465–505. Marcel Dekker, New York.

DWASS, M. (1957): “Modified Randomization Tests for Nonparametric Hypotheses,” *Annals of Mathematical Statistics*, 28, 181–187.

EFRON, B. (1982): *The Jacknife, the Bootstrap and Other Resampling Plans*, CBS-NSF Regional Conference Series in Applied Mathematics, Monograph No. 38. Society for Industrial and Applied Mathematics, Philadelphia, PA.

EFRON, B., AND R. J. TIBSHIRANI (1993): *An Introduction to the Bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York.

ENGLE, R. F. (1982): “Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation,” *Econometrica*, 50, 987–1007.

FAUST, J., AND E. M. LEEPER (1997): “When Do Long-Run Identifying Restrictions Give Reliable Results,” *Journal of Business and Economic Statistics*, 15, 345–353.

FIELLER, E. C. (1940): “The Biological Standardization of Insulin,” *Journal of the Royal Statistical Society (Supplement)*, 7, 1–64.

——— (1954): “Some Problems in Interval Estimation,” *Journal of the Royal Statistical Society, Series B*, 16, 175–185.

FISHER, R. A. (1932): *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

FOLKS, J. L. (1984): "Combination of Independent Tests," in *Handbook of Statistics, Volume 4, Nonparametric Methods*, ed. by P. R. Krishnaiah, and P. K. Sen, pp. 113–121. North-Holland, Amsterdam.

GALLANT, A. R., AND G. TAUCHEN (1996): "Which Moments to Match?," *Econometric Theory*, 12, 657 – 681.

GIBBONS, M. R., S. A. ROSS, AND J. SHANKEN (1989): "A Test of the Efficiency of a Given Portfolio," *Econometrica*, 57, 1121–1152.

GOFFE, W. L., G. D. FERRIER, AND J. ROGERS (1994): "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, 60, 65–99.

GOURIÉROUX, C., AND A. MONFORT (1996): *Simulation-Based Econometric Methods*. Oxford University Press, Oxford, U.K.

HAJIVASSILIOU, V. A. (1993): "Simulation Estimation Methods for Limited Dependent Variables," in Maddala, Rao, and Vinod (1993), pp. 519–543.

HALL, P. (1992): *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

HARVEY, A. C., AND G. D. A. PHILLIPS (1980): "Testing for Serial Correlation in Simultaneous Equation Models," *Econometrica*, 48, 747–759.

——— (1982): "Testing for Contemporaneous Correlation of Disturbances in Systems of Regression Equations," *Bulletin of Economic Research*, 34(2), 79–81.

HOROWITZ, J. L. (1997): "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in *Advances in Economics and Econometrics*, ed. by D. Kreps, and K. W. Wallis, vol. 3, pp. 188–222. Cambridge University Press, Cambridge, U.K.

JARQUE, C. M., AND A. K. BERA (1980): "Efficient Tests for Normality, Heteroscedasticity and Serial Independence of Regression Residuals," *Economics Letters*, 6, 255–259.

JARQUE, C. M., AND A. K. BERA (1987): "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, 55, 163–172.

JEONG, J., AND G. S. MADDALA (1993): "A Perspective on Application of Bootstrap Methods in Econometrics," in Maddala, Rao, and Vinod (1993), pp. 573–610.

KEANE, M. P. (1993): "Simulation Estimation for Panel Data Models with Limited Dependent Variables," in Maddala, Rao, and Vinod (1993), pp. 545–571.

KIVIET, J. F., AND J.-M. DUFOUR (1997): "Exact Tests in Single Equation Autoregressive Distributed Lag Models," *Journal of Econometrics*, 80, 325–353.

KOLMOGOROV, A. N. (1933): "Sulla determinazione empiricadi una legge di distribuzione," *Giorna. Ist. Attuari.*, 4, 83–91.

LAITINEN, K. (1978): "Why is Demand Homogeneity so Often Rejected?," *Economics Letters*, 1, 187–191.

LEE, J. H., AND M. L. KING (1993): "A Locally Most Mean Powerful Based Score Test for ARCH and GARCH Regression Disturbances," *Journal of Business and Economic Statistics*, 11, 17–27, Correction 12 (1994), 139.

LEE, J. H. H. (1991): "A Lagrange Multiplier Test for GARCH Models," *Economics Letters*, 37, 265–271.

MADDALA, G. S., C. R. RAO, AND H. D. VINOD (eds.) (1993): *Handbook of Statistics 11: Econometrics*. North-Holland, Amsterdam.

MARIANO, R. S., AND B. W. BROWN (1993): "Stochastic Simulation for Inference in Nonlinear Errors-in-Variables Models," in Maddala, Rao, and Vinod (1993), pp. 611–627.

MCCLOSKEY, D. N., AND S. T. ZILIAK (1996): "The Standard Error of Regressions," *The Journal of Economic Literature*, XXXIV, 97–114.

NELSON, C. R., AND R. STARTZ (1990a): "The Distribution of the Instrumental Variable Estimator and its t-ratio When the Instrument is a Poor One," *Journal of Business*, 63, 125–140.

——— (1990b): "Some Further Results on the Exact Small Properties of the Instrumental Variable Estimator," *Econometrica*, 58, 967–976.

NELSON, C. R., R. STARTZ, AND E. ZIVOT (1996): "Valid Confidence Intervals and Inference in the Presence of Weak Instruments," Discussion paper, Department of Economics, University of Washington.

NEYMAN, J., AND E. S. PEARSON (1933): "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical transactions of the Royal society A*, 231, 289–337.

NIELSEN, B. (1998): "On the Distribution of Tests for the Cointegration Rank," Discussion paper, Nuffield College, Oxford, U.K.

PEARSON, K. (1933): "On a Method of Determining Whether a Sample of Size n Supposed to Have Been Drawn from a Parent Population," *Biometrika*, 25, 379–410.

PHILLIPS, P. C. B. (1983): "Exact Small Sample theory in the Simultaneous Equations Model," in *Handbook of Econometrics, Volume 1*, ed. by Z. Griliches, and M. D. Intrilligator, chap. 8, pp. 449–516. North-Holland, Amsterdam.

SAPHORES, J.-D., L. KHALAF, AND D. PELLETIER (1998): "Modelling Unexpected Changes in Stumpage Prices: An Application to Pacific Northwest National Forests," Discussion paper, GREEN, Université Laval, Québec.

SAVIN, N. E. (1984): "Multiple Hypothesis Testing," in *Handbook of Econometrics, Volume 2*, ed. by Z. Griliches, and M. D. Intrilligator, pp. 827–879. North-Holland, Amsterdam.

SAVIN, N. E., AND A. H. WÜRTZ (1998): "The Effect of Nuisance Parameters on Critical Values and Power: Lagrange Multiplier Tests in Logit Models," Discussion paper, Department of Economics, University of Iowa and Department of Economics, University of Aarhus.

SHAO, S., AND D. TU (1995): *The Jackknife and Bootstrap*. Springer-Verlag, New York.

SMIRNOV, N. V. (1939): "Sur les écarts de la courbe de distribution empirique (Russian/French Summary)," *Matematičeskiĭ Sbornik N.S.*, 6, 3–26.

STAIGER, D., AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

STEWART, K. G. (1997): "Exact Testing in Multivariate Regression," *Econometric Reviews*, 16, 321–352.

STOCK, J. H., AND J. WRIGHT (1997): "GMM and Weak Identification," Discussion paper, Kennedy School of Government, Harvard University, Forthcoming in Econometrica.

TIPPETT, L. H. (1931): *The Methods of Statistics*. Williams and Norgate, London.

VINOD, H. D. (1993): "Bootstrap Methods: Applications in Econometrics," in Maddala, Rao, and Vinod (1993), pp. 629–661.

WANG, J., AND E. ZIVOT (1996): "Inference on a Structural Parameters in Instrumental Variable Regression with Weakly Correlated Instruments," Discussion paper, Department of Economics, University of Washington.

WILKINSON, B. (1951): "A Statistical Consideration in Psychological Research," *Psychology Bulletin*, 48, 156–158.

WILKS, S. S. (1932): "Certain Generalizations in the Analysis of Variance," *Biometrika*, 24, 471–494.

ZELLNER, A. (1962): "An Efficient Method for Estimating Seemingly Unrelated Regressions and Tests for Aggregate Bias," *Journal of the American Statistical Association*, 57, 348–368.