Monte Carlo Test Methods in Econometrics *

Jean-Marie Dufour † and Lynda Khalaf ‡

First version: October 1998
Revised: February 1999
This version: August 1999
Compiled: 11th April 2001, 10:52pm

This paper has been published in: *Companion to Theoretical Econometrics*, edited by Badi Baltagi, Blackwell, Oxford, U.K., 2001, Chapter 23, 494-519.

* The authors thank three anonymous referees and the Editor Badi Baltagi for several useful comments. 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 Social Sciences and Humanities Research Council of Canada, the Natural Sciences and Engeneering Council of Canada, and the Government of Québec (Fonds FCAR).

† 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

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, Monfort, and Renault (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 often raises 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 discussed 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, and 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 include:

- 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.

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 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 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: 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)
  - ▶ MC tests or Bartlett Corrections?
  - ▶ The Case of Unidentified Nuisance Parameters (test for structural jumps, test for ARCH-M)
- 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

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 $\theta \in \Theta$, where Θ refers to the parameter space (usually a vector space), $\mathcal{Y}$ is the sample space, P_θ is the family of probability distributions on $\mathcal{Y}$. Furthermore, let Y denote the observations, and Θ_0 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 acceptance region), and its complement whose elements are viewed as inconsistent with H_0 (the rejection region, 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.

Without loss of generality, we suppose the critical region has the form $S(Y) \geq c$. To obtain a test of level α, c must be chosen so that the probability of rejecting the null hypothesis $\mathsf{P}_\theta[S(Y) \geq c]$ when H_0 is true (the probability of a *type I error*) is not greater than α, *i.e.* we must have:

$$\sup_{\theta \in \Theta_0} \mathsf{P}_\theta[S(Y) \geq c] \leq \alpha \,. \tag{2.1}$$

Further, the test has *size* α if and only if

$$\sup_{\theta \in \Theta_0} \mathsf{P}_\theta[S(Y) \geq c] = \alpha \,. \tag{2.2}$$

To solve for c in (2.1) or (2.2), 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 (2.1) - (2.2) that, in many cases of practical interest, the distribution of $S(Y)$ may be different for different parameter values. When the null hypothesis completely fixes the value of θ (*i.e.* Θ_0 is a point), the hypothesis is called a *simple hypothesis*. Most hypotheses encountered in practice are **composite,** *i.e.* the set Θ_0 contains 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 are called **nuisance parameters.**

When we talk about an **exact test,** it must be understood that attention is restricted to level-correct critical regions, where (2.1) must hold for a given finite sample size, for all values of the parameter θ compatible with the null. Consequently, in carrying out an exact test, one may encounter two problems. The first one is to extract the analytic form of the distribution of $S(Y)$. The second one is to maximize the rejection probability over the relevant nuisance parameter space, subject to the level constraint. We will see below that the first problem can easily be solved when Monte Carlo test techniques are applicable. The second one is usually more difficult to tackle, and its importance is not fully recognized in econometric practice.

A reasonable solution to both problems often exists when one is dealing with large samples. Whereas the null distribution of $S(Y)$ may be complicated and/or may involve unknown parameters, its asymptotic null distribution in many common cases has a known form and is nuisance-parameter-free (*e.g.*, a normal or chi-square distribution). 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 critical values which yield the desired size α for a given sample size can be very different from these approximate values obtained through an asymptotic argument. 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 empirically relevant sample sizes. The problem, and our main point, is that **finite sample accuracy is not merely a small sample problem.**

2.1. Instrumental regressions

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

$$\begin{aligned} y &= Y\beta + X_1\gamma_1 + u = Z\delta + u\,, && (2.3)\\ Y &= X_1\Pi_1 + X_2\Pi_2 + V\,, && (2.4) \end{aligned}$$

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:

$$\begin{aligned} \left[\begin{array}{cc} y & Y \end{array}\right] &= X\Pi + \left[\begin{array}{cc} v & V \end{array}\right],\ \Pi = \left[\begin{array}{cc} \pi_1 & \Pi_1 \\ \pi_2 & \Pi_2 \end{array}\right], && (2.5)\\ \pi_1 &= \Pi_1\beta + \gamma_1\,,\ \ \pi_2 = \Pi_2\beta\,. && (2.6) \end{aligned}$$

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

$$rank(\Pi_2) = m\,. \tag{2.7}$$

To test the general linear hypothesis $R\delta = r$, where R is a full row rank $q \times (m + k)$ matrix, 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

$$\hat{\delta} = [Z'P(P'P)^{-1}P'Z]^{-1}Z'P(P'P)^{-1}P'y \tag{2.8}$$

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

$$\tau_w = \frac{1}{s^2}(r - R\hat{\delta})'[R'(ZP(P'P)^{-1}P'Z)^{-1}R]\ (r - R\hat{\delta}) \tag{2.9}$$

where $s^2 = \frac{1}{n}(y - Z\hat{\delta})'(y - Z\hat{\delta})'$. Under usual regularity conditions and imposing identification, τ_w is distributed like a $\chi^2(q)$ variable, where $q = rank(R)$.

Bartlett (1948) and Anderson and Rubin (1949, henceforth 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 $F-$ statistic for testing whether the coefficients of the regressors X_2 "excluded" from (2.3) regressors are zero in the regression of y^* on all the exogenous variables, *i.e.* we simply test $\gamma_2 = 0$ in the extended linear regression $y^* = X_1\gamma_1 + X_2\gamma_2 + u$.

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 (2.3) with two endogenous variables ($p = 2$) and $k = 1$ exogenous variables. The structural equation includes only the endogenous variable. The restrictions tested are of the form $H_{01} : \beta = \beta^0$. The sample sizes are set to n = 25, 100, 250. The exogenous regressors are independently drawn from the standard normal distribution. These are

Table 2.1: IV-based Wald /Anderson-Rubin tests: empirical type I errors

Π_2	$n = 25$		$n = 100$		$n = 250$	
	Wald	AR	Wald	AR	Wald	AR
1	.061	.059	.046	.046	.049	.057
.9	.063	.059	.045	.046	.049	.057
.7	.071	.059	.046	.046	.052	.057
.5	.081	.059	.060	.046	.049	.057
.2	.160	.059	.106	.046	.076	.057
.1	.260	.059	.168	.046	.121	.057
.05	.332	.059	.284	.046	.203	.057
.01	.359	.059	.389	.046	.419	.057

drawn only once. The errors are generated according to a multinormal distribution with mean zero and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & .95 \\ .95 & 1 \end{bmatrix} . \tag{2.10}$$

The other coefficients are:

$$\beta = \beta^0 = 0\,;\, \Pi_2 = 1\,,\ .9\,,\ .7\,,\ .5\,,\ .2\,,\ .1\,,\ .05\,,\ .01. \tag{2.11}$$

In this case, the 2SLS-based test corresponds to the standard t-test [see Nelson and Startz (1990b) for the relevant formulae]. 1000 replications are performed. Table 2.1 reports probabilities of type I error [$P(type\, I\, error)$] associated with the two-tailed 2SLS t-test for the significance of β and the corresponding Anderson-Rubin 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.[1] It is evident that IV-based Wald tests perform very poorly in terms of size control. Identification problems severely distort test sizes. While the evidence of size distortions is notable even in identified models, the problem is far more severe in near-unidentified situations. 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 sizes. 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 (1996) 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. On general linear structural restrictions, see Dufour and Khalaf (1998b).

Despite the recognition of the need for caution in the application of IV-based tests, 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 Wald tests realize computational savings at the risk of very poor reliability.

[1]The problem is more complicated when the structural equation includes more than one endogenous variable. See Dufour and Khalaf (1998b) for a detailed discussion of this case.

2.2. Normality tests

Let us now consider the fundamental problem of testing disturbance normality in the context of the linear regression model:

$$Y = X\beta + u \tag{2.12}$$

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 $u = (u_1, \ldots, u_n)'$ is an n-dimensional vector of $i.i.d$ disturbances. The problem consists in testing:

$$H_0 : f(x) = \varphi(x;\, 0, \sigma)\ ,\ \sigma > 0\,, \tag{2.13}$$

where $f(x)$ is the probability density function (pdf) of u_i, and $\varphi(x;\, \mu, \sigma)$ 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

$$\widehat{u} = y - X\widehat{\beta} = M_X u \tag{2.14}$$

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

$$s^2 = (n-k)^{-1} \sum_{i=1}^{n} \widehat{u}_{in}^2\ ,\quad \widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \widehat{u}_{in}^2\ . \tag{2.15}$$

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.

The KS test is based on a measure of discrepancy between the empirical and hypothesized distributions:

$$KS = \max\left(D^+,\, D^-\right) \tag{2.16}$$

where $D^+ = \max_{1 \leq i \leq n}[(i/n) - \widehat{z}_i]$ and $D^- = \max_{1 \leq i \leq n}[\widehat{z}_i - (i-1)/n]$, $\widehat{z}_i = \Phi(\widehat{u}_{in}/s)$, $i = 1, \ldots, n$, and $\Phi(.)$ denotes the cumulative $N(0,1)$ distribution function. The exact and limiting distributions of the KS statistic are non-standard and even asymptotic points must be estimated. We have used significance points from D'Agostino and Stephens (1986, Table 4.7), although these were formally derived for the location-scale model. The JB test combines the skewness (Sk) and kurtosis (ku) coefficients:

$$JB = n\left[\frac{1}{6}(Sk)^2 + \frac{1}{24}(Ku-3)^2\right] \tag{2.17}$$

where $Sk = n^{-1}\sum_{i=1}^{n} \widehat{u}_{in}^3/(\widehat{\sigma}^2)^{3/2}$ and $Ku = n^{-1}\sum_{i=1}^{n} \widehat{u}_{in}^4/(\widehat{\sigma}^2)^2$. Under the null and appropriate regularity conditions, the JB statistic is asymptotically distributed as $\chi^2(2)$; the statistic's exact distribution is intractable.

We next summarize relevant results from the simulation experiment reported in Dufour, Farhat, Gardiol, and Khalaf (1998). The experiment based on (2.12) was performed as follows. 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. Table 2.2 reports rejection percentages (from 10000 replications) at the nominal size of 5% under the null hypothesis, with n = 25, 50, 100,

Table 2.2: Kolmogorov-Smirnov / Jarque-Bera residuals based tests: empirical type I errors

k_1		$n = 25$		$n = 50$		$n = 100$	
		KS	JB	KS	JB	KS	JB
0	STD	.050	.029	.055	.039	.055	.041
	MC	.052	.052	.052	.050	.047	.048
$2, \quad (n = 25)$	STD	.114	04.8	.163	.064	.131	.131
$4, \quad (n > 25)$	MC	.053	05.2	.050	.050	.050	.050
$k, \quad (n \leq 50)$	STD	.282	06.7	.301	.084	.322	.322
$8, \quad (n = 100)$	MC	.052	04.8	.050	.047	.047	.047

Note: STD refers to the standard normality test and MC denotes the (corresponding) Monte Carlo test.

$k =$ the largest integer less than or equal to $\sqrt{n}$ and $k_1 = 0, 2, 4, ..., k-1$. 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. We will discuss the MC tests results in Section 4.

2.3. Uniform linear hypothesis in multivariate regression models

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:

$$Y = XB + U \tag{2.18}$$

where $Y = [Y_1, ..., Y_p]$ 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 $\det(\Sigma) \neq 0$. To derive the distribution of the relevant test statistics, we also assume the following:

$$\widetilde{U}_i = JW_i, \; i = 1, \ldots, n, \tag{2.19}$$

where the vector $w = vec(W_1, \ldots, W_n)$ has a known distribution and J is an unknown, non-singular matrix; for further reference, let $W = [W_1, \ldots, W_n]' = UG'$, where $G = J^{-1}$. In particular, this condition will be satisfied when the normality assumption is imposed. An alternative representation of the model is

$$Y_{ij} = \alpha_j + \sum_{k=1}^{p} \beta_{jk} X_{ik} \;, \; i = 1 \;, \; ... \;, \; n, \;\; j = 1 \;, \; ... \;, \; p \;. \tag{2.20}$$

Uniform Linear (UL) constraints take the special form

$$H_0 : RBC = D \tag{2.21}$$

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

$$H_{01} : R\beta_i = \delta_i \, , \; i = 1, \, \dots \, , \, p \, , \tag{2.22}$$

which corresponds to $C = I_p$. Here we shall focus on hypotheses of the form (2.22) for ease of exposition; see Dufour and Khalaf (1998c) for the general case.

Stewart (1997) discusses several econometric applications where the problem can be stated in terms of UL hypotheses. A prominent example includes the multivariate test of the capital asset pricing model (CAPM). Let $r_{jt}, \; j = 1, \, \dots \, , \, p$, 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 CAPM model:

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

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. The latter hypothesis is a special case of (2.22) where 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 (2.20) where the hypothesis of linear homogeneity corresponds to

$$H_0 : \sum_{k=1}^{p} \beta_{jk} = 0 \, , \; j = 1 \, , \, ... \, , \, p \, . \tag{2.23}$$

In this context, the likelihood ratio (LR) criterion is:

$$LR = n \, \ln(\Lambda) \, , \; \Lambda = |\widehat{U}_0'\widehat{U}_0| / |\widehat{U}'\widehat{U}| \tag{2.24}$$

where $\widehat{U}_0'\widehat{U}_0$ and $\widehat{U}'\widehat{U}$ are respectively the constrained and unconstrained SSE matrices. On observing that, under the null hypothesis,

$$\widehat{U}'\widehat{U} \;=\; G^{-1}W'MW \left(G^{-1}\right)' , \tag{2.25}$$

$$\widehat{U}_0'\widehat{U}_0 \;=\; G^{-1}W'M_0W'(G^{-1})', \tag{2.26}$$

where $M_0 = I - X(X'X)^{-1}(X'X - R'(R(X'X)^{-1}R')^{-1}R)(X'X)^{-1}X'$ and $M = I - X(X'X)^{-1}X'$, we can then rewrite Λ in the form

$$\Lambda = |W'M_0W| / |W'MW| \tag{2.27}$$

where the matrix $W = UG'$ has a distribution which does not involve nuisance parameters. As shown in Dufour and Khalaf (1998c), decomposition (2.26) obtains only in the case of UL constraints. In Section 4 we will exploit the latter result to obtain exact MC tests based on the LR statistic.

To illustrate the performance of the various relevant tests, we consider a simulation experiment modelled after demand homogeneity tests, *i.e.* (2.20) and (2.23) with $p = 5, 7, 8$, $n = 20 \, , 25 \, , 40 \, , 50 \, , 100$. 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.

Table 2.3: Empirical type I errors of multivariate tests: uniform linear hypotheses

Sample Size	5 Equations			7 Equations			8 Equations		
	LR	LR_c	LR_{MC}	LR	LR_c	LR_{MC}	LR	LR_c	LR_{MC}
20	.295	.100	.051	.599	.250	.047	.760	.404	.046
25	.174	.075	.049	.384	.145	.036	.492	.190	.042
40	.130	.066	.056	.191	.068	.051	.230	.087	.051
50	.097	.058	.055	.138	.066	.050	.191	.073	.053
100	.070	.052	.042	.078	.051	.041	.096	.052	.049

Note: LR, LR_c, LR_{MC} denote (respectively) the standard LR test, the Bartlett corrected test and the (corresponding) MC test.

The coefficients for all experiments are available from Dufour and Khalaf (1998c). The statistics examined are the relevant LR criteria defined by (2.24) and the Bartlett-corrected LR test [Attfield (1995, section 3.3)]. The results are summarized in Table 2.3. We report the tests empirical size, based on a nominal size of 5% and 1000 replications. It is evident that the asymptotic LR test overrejects 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. We will discuss the MC tests results in Section 4.

To conclude this section, it is worth noting that an exact test is available for hypotheses of the form $H_0 : RBC = D$, where $\min(r, c) \leq 2$. 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.[1]

2.4. Econometric applications: discussion

In many empirical problems, it is quite possible that the exact null distribution of the relevant test statistic $S(Y)$ will not be easy to compute analytically, even though it is nuisance-parameter-free. 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 and the UL hypotheses tests, all involve pivotal statistics. The problem is more complicated in the presence of nuisance parameters. We will first discuss a property related to nuisance-parameter-dependent test statistics which will prove to be fundamental in finite sample contexts.[2]

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

$$S(Y) \leq S^*(Y), \quad \forall \theta \in \Theta_0 , \tag{2.28}$$

and $S^*(Y)$ is pivotal under the null. Then $S(Y)$ is said to be **boundedly pivotal.** The implications of

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

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

this property are as follows. From (2.28), 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

$$\mathsf{P}[S^*(Y) \geq c] = \alpha\,, \tag{2.29}$$

we solve the level constraint for the test based on $S(Y)$. It is clear that (2.28) and (2.29) 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 discussed in Sections 2.1 and 2.3 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! For further discussion of the basic notions of statistical testing mentioned in this section, the reader may consult Lehmann (1986, Chapter 3), Gouriéroux and Monfort (1995) and Dufour (1990, 1997) .

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 $X_1,\ \ldots,\ X_m$ and $Y_1,\ \ldots,\ Y_n$, where $X_1,\ \ldots,\ X_m \overset{iid}{\sim} F(x)$, $Y_1,\ \ldots,\ Y_n \overset{iid}{\sim} F(x-\delta)$ 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 $z = (X_1,\ \ldots,\ X_m,\ Y_1,\ \ldots,\ Y_n)$ and $s = \frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- Obtain all possible $Q = (n+m)!$ permutations of z, $z^{(1)},\ \ldots,\ z^{(Q)}$, and calculate the associated "permuted analogues" of s

$$s^{(j)} = \frac{1}{m}\sum_{i=1}^{m} z_i^{(j)} - \frac{1}{n}\sum_{i=m+1}^{m+n} z_i^{(j)}\ , \quad j = 1,\ \ldots\ ,\ Q\,.$$

- 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 we wish to get a $5\%-$level permutation test, and 99 random permutations can be generated, 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 that Barnard (1963) proposed later a similar idea.[1]

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 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

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

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

[1] Bera and Jarque (1982), Breusch and Pagan (1979, 1980) have also proposed related simulation-based techniques. However, these authors do not provide finite-sample theoretical justification for the proposed procedures. In particular, in contrast with Dwass (1957) and Barnard (1963) (and similarly to many other later authors who have proposed exploiting Monte carlo techniques), they do not observe that appropriately randomized tests allow one to exactly control the level of a test in finite samples.

gives the rank of S_0 among the variables $S_0, S_1, \ldots, S_N$.[1] Then the critical region of a test with level α is:

$$\widehat{p}_N(S_0) \leq \alpha \tag{3.1}$$

where , $0 < \alpha < 1$ and

$$\widehat{p}_N(x) = \frac{N\widehat{G}_N(x) + 1}{N + 1} \,. \tag{3.2}$$

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 a 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

$$\widehat{R}_N(S_0) \geq (N + 1)(1 - \alpha) + 1 \,. \tag{3.3}$$

In other words, for 99 replications a 5% MC test is significant if the rank of S_0 in the series $S_0, S_1, \ldots, S_N$ 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 (3.1) 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.

Theorem 3.1.1 *Consider an $M \times 1$ vector of exchangeable real random variables $(Y_1, \ldots, Y_M)$ 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 $Y_1, \ldots, Y_M$. Then*

$$\mathsf{P}\left[\tfrac{R_j}{M} \geq z\right] = \tfrac{I[(1-z)M]+1}{M}, \quad 0 < z \leq 1 \,. \tag{3.4}$$

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

If S is a 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

[1]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 $S_0, S_1, \ldots, S_N$. Throughout this section N refers to the number of MC replications.

(3.4), 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 asymptotic theory has 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,

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

in the context of the linear regression model $Y = X\beta + u$.[1] 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 $JB^{(0)}$.
- Treating 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 $JB^{(j)}$ associated with the regression of $\widetilde{Y}^{(j)}$ on X.
- Obtain the rank $\widehat{R}_N(JB^{(0)})$ in the series $JB^{(0)}, JB^{(1)}, \ldots, JB^{(N)}$..
- 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 $\widehat{p}_N(S_0) = [N+1-\widehat{R}_N(S_0)]/(N+1)$. Dufour, Farhat, Gardiol, and Khalaf (1998) show that the JB statistics can be computed from the standardized residual vector $\widehat{u}/s$. Using (2.14), we see that

$$\widehat{u}/s = \frac{\widehat{u}}{(\widehat{u}'\widehat{u}/(n-k))^{1/2}} = (n-k)^{1/2}\frac{M_X u}{(u'M_X u)^{1/2}} = (n-k)^{1/2}\frac{M_X w}{(w'M_X w)^{1/2}}\ , \tag{3.5}$$

where $w = u/\sigma \overset{i.i.d.}{\sim} N(0,1)$ when $u \sim N(0, \sigma^2 I_n)$. It follows that the simulated statistics $JB^{(j)}$ may be obtained using draws from a nuisance-parameter free (standard normal) null distribution.

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.

[1] See Section 2.2 for a formal presentation of the model and test statistics. Some equations are redefined here for convenience.

- 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 (3.2).

- 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,

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

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), it is shown 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)]. SA starts from an initial point (it is natural to use $\widehat{\theta}_n^0$ here) 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.[1]

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) entails 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. 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: normality tests, heteroskedasticity tests including tests for (G)ARCH

[1]Global optimization is generally considered to be (relatively) computationally demanding. We have experimented (see Dufour and Khalaf (1998c, 1998b)) with several MMC tests where the number of nuisance parameters referred to the Simulated Annealing algorithm was up to 20. Our simulations show that the method works well. Convergence was slow in some cases (less than 5 per 1000). Recall however that for the problem at hand, one just practically needs to check whether the maximized function exceeds α, which clearly reduces the computational burdens.

and tests for break in variance at unknown points, independence tests and tests based on autocorrelations.[1] 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; power studies are also reported on a case by case basis to assess the performance of MC size corrected tests.

Relevant results pertaining to the examples considered above are included in Tables 2.2 - 2.3. It is evident from Table 2.2 that the size of the JB and KS tests is perfectly controlled for all designs considered.[2] Table 2.3 includes the empirical size of the MC LR test for linear restrictions. From (2.27), we see that under the distributional assumption (2.19), the simulated statistics may be obtained using draws from a nuisance-parameter free null distribution, namely the hypothesized distribution of the vector w. Consequently, application of the MC test procedure yields exact p-values. Indeed, it is shown in Table 2.3 that the MC LR test achieves perfect size control.[3]

Now to illustrate the feasibility of MMC tests and the usefulness of BMC tests, we will focus on 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 provide examples 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. Consider the three equations system

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11} X_1 + U_1 \ , \\ Y_2 &= \beta_{20} + \beta_{22} X_2 + U_2 \ , \\ Y_3 &= \beta_{30} + \beta_{33} X_3 + U_3 \ , \end{aligned} \tag{4.1}$$

imposing normality, and the hypothesis $H_0 : \beta_{11} = \beta_{22} = \beta_{33}$. First restate H_0 in terms of the MLR model which includes the SURE system as a special case, so that it incorporates the SURE exclusion restrictions. Formally, in the framework of the MLR model

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11} X_1 + \beta_{12} X_2 + \beta_{13} X_3 + U_1 \ , \\ Y_2 &= \beta_{20} + \beta_{21} X_1 + \beta_{22} X_2 + \beta_{23} X_3 + U_2 \ , \\ Y_3 &= \beta_{30} + \beta_{31} X_1 + \beta_{32} X_2 + \beta_{33} X_3 + U_3 \ , \end{aligned} \tag{4.2}$$

H_0 is equivalent to the joint hypothesis

$$H_0^* : \beta_{11} = \beta_{22} = \beta_{33} \text{ and } \beta_{12} = \beta_{13} = \beta_{21} = \beta_{23} = \beta_{31} = \beta_{32} = 0 \, . \tag{4.3}$$

The associated LR statistic is

$$LR = n \ln(\Lambda) \ , \quad \Lambda = |\widehat{\Sigma}_0| / |\widehat{\Sigma}| \tag{4.4}$$

[1] 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.

[2] See Dufour, Farhat, Gardiol, and Khalaf (1998) for the power study.

[3] See Dufour and Khalaf (1998c) for the power study.

where $\widehat{\Sigma}_0$ and $\widehat{\Sigma}$ are the restricted and unrestricted SURE MLE. We also consider

$$LR^* = n \ln(\Lambda^*), \quad \Lambda^* = |\widehat{\Sigma}_0| / |\widehat{\Sigma}_u| \tag{4.5}$$

where $\widehat{\Sigma}_u$ is the unconstrained estimate of Σ in the "nesting" MLR model. Since the restricted model is the same in both LR and LR^*, while the unrestricted model used in LR^* includes as a special case the one in LR, it is straightforward to see that $LR \le LR^*$, so that the distribution function of LR^* provides an upper bound on the distribution function of LR.

In order to apply a BMC test, we need to construct a set of UL restrictions that satisfy (4.3) so that the corresponding LR^* criterion conforming with these UL restrictions yields a valid bound on the distribution of LR. Indeed, as emphasized above, the LR test statistic for UL restrictions is pivotal. Furthermore, by considering UL restrictions obtained as a special case of H_0^*, we can be sure that the associated statistic is always $\ge LR$. Here, it is easy to see that the constraints setting the coefficients β_{ij}, i, $j = 1, \ldots, 3$, to specific values meet this criterion. Note that the statistic just derived serves to bound both LR and LR^*.

Define $\theta \equiv C(\Sigma)$ as the vector of the parameters on or below the diagonal of the Cholesky factor $T(\Sigma)$ of the covariance matrix Σ [i.e., $T(\Sigma)$ is the lower triangular matrix such that $T(\Sigma)T(\Sigma)' = \Sigma$]. 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: $U^{(1)}, U^{(2)}, \ldots, U^{(p)}$ 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 bound MC procedure $BMC(\theta)$, described below, for $\theta \equiv C(\widehat{\Sigma}_0)$. The Cholesky decomposition is used to impose positive definiteness and avoid redundant parameters. The output is the BMC p-value. Reject the null if the latter is $\le .05$ and STOP.
- Otherwise, call the procedure $MC(\theta)$, also described below, for $\theta \equiv C(\widehat{\Sigma}_0)$. It is important to note here that Σ is the only relevant nuisance parameter, for the example considered involves linear constraints [see Breusch (1980)]. 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 example, Simulated Annealing) for the function $MC(\theta)$ using $\theta \equiv C(\widehat{\Sigma}_0)$ 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 $MC(\theta)$:

- 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, $U^{(j)}$, 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.
- The $BMC(\theta)$ procedure may be obtained as just described, replacing $LR^{*(j)}$by $LR_c^{(j)}$. Alternatively, the BMC procedure may be rewritten following the methodology relating to MC tests of UL hypotheses so that no (unknown) parameters intervene in the generation of the simulated (bounding) statistics. Indeed, the bounding statistic satisfies (2.27) under (2.19). Thus $LR_c^{(j)}$ may be obtained using draws from, *e.g.*, the multivariate independent normal distribution.

In Dufour and Khalaf (1998c), 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 show that the MC test procedure achieves perfect size control and has good power. The same methodology may also be applied in simultaneous equations models such as (2.3). In Dufour and Khalaf (1998b), we present simulations which illustrate the performance of limited-information LR-based MC tests in this context. We have attempted to apply the MC test procedure to the IV-based Wald-type test for linear restrictions on structural parameters. 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 (refer to Section 2.3) 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)]

These results are also related to the non-invariance problems associated with Wald tests in nonlinear contexts [see, *e.g.,* Dufour (1997) and Dagenais and Dufour (1991)]. Indeed, it is evident from (2.3)-(2.5) that *seemingly linear* constraints on structural coefficients in instrumental regressions often involve non-linear hypotheses implied by the structure. Of course, not all Wald tests will suffer from such problems. For instance, Wald tests for linear restrictions in linear regression models yield exact F-tests.

We conclude this section with a specific problem where the MC test strategy conveniently solves a difficult and non-standard distributional problems: the problem of unidentified nuisance parameters.

4.3. Non-identified nuisance parameters

The example we discuss here 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:

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

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 $\left[\ t-1,\quad t\ \right]$; 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. Bernard, Dufour, Khalaf, and Genest (1998) consider another illustrative example: testing for ARCH-in-mean effects, and show that the MC method works very well in terms of size and power.

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.

Of course, the first purpose of the MC test technique is to control the probability of type I errors (below a given *level*) so that rejections can properly be interpreted as showing that the null hypothesis is "incompatible" with the data. However, once level is controlled, we can (and should) devote more attention to finding procedures with good *power* properties. Indeed, by helping to put the problem of level control out of the way, we think the technique of MC tests should help econometricians devote research to power issues as opposed to level. So an indirect consequence of the implementation of the technique may well be an increased emphasis on the design of more powerful tests.

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

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.

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

BERA, A. K., AND C. M. JARQUE (1982): "Model Specification Tests: A Simultaneous Approach," *Journal of Econometrics*, 20, 59–82.

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.

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. (1980): "Useful Invariance Results for Generalized Regression Models," *Journal of Econometrics*, 13, 327–340.

BREUSCH, T. S., AND A. R. PAGAN (1979): "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, 47, 1287–1294.

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.

DAGENAIS, M. G., AND J.-M. DUFOUR (1991): "Invariance, Nonlinear Models and Asymptotic Tests," *Econometrica*, 59, 1601–1615.

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

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 (1996): “Finite Sample Inference Methods for Simultaneous Equations and Models with Unobserved and Generated Regressors,” Discussion paper, C.R.D.E., Université de Montréal.

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.

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.

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 (1995): *Statistics and Econometric Models, Volumes One and Two*. Cambridge University Press, Cambridge, U.K.

——— (1996): *Simulation-Based Econometric Methods*. Oxford University Press, Oxford, U.K.

GOURIÉROUX, C., A. MONFORT, AND E. RENAULT (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8S, 85–118.

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.

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.

LEHMANN, E. L. (1986): *Testing Statistical Hypotheses, 2nd edition*. John Wiley & Sons, New York.

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.

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.

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.

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.

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

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.