Foundations and Trends® in
Econometrics
Vol. 5, Nos. 1–2 (2011) 1–181

DOI: 10.1561/0800000011

now
the essence of knowledge

Monte Carlo Simulation for Econometricians

By Jan F. Kiviet

Contents

Foundations and Trends® in
Econometrics
Vol. 5, Nos. 1–2 (2011) 1–181

DOI: 10.1561/0800000011

now
the essence of knowledge

Monte Carlo Simulation for Econometricians

Jan F. Kiviet[1,2]

[1] *Division of Economics, School of Humanities and Social Sciences, Nanyang Technological University, Singapore 637332, jfkiviet@ntu.edu.sg*
[2] *Department of Quantitative Economics, Amsterdam School of Economics, University of Amsterdam, 1018 XE Amsterdam, The Netherlands, j.f.kiviet@uva.nl*

Abstract

Many studies in econometric theory are supplemented by Monte Carlo simulation investigations. These illustrate the properties of alternative inference techniques when applied to samples drawn from mostly entirely synthetic data generating processes. They should provide information on how techniques, which may be sound asymptotically, perform in finite samples and then unveil the effects of model characteristics too complex to analyze analytically. Also the interpretation of applied studies should often benefit when supplemented by a dedicated simulation study, based on a design inspired by the postulated actual empirical data generating process, which would come close to bootstrapping. This review presents and illustrates the fundamentals of conceiving and executing such simulation studies, especially synthetic but also more dedicated, focussing on controlling their accuracy, increasing their efficiency, recognizing their limitations, presenting their results in a coherent and palatable way, and on the appropriate interpretation of their actual findings, especially when the simulation study is used to rank the qualities of alternative inference techniques.

Preface and Overview

Since many decades much of the research in econometric theory is supported or illustrated by Monte Carlo simulation studies. Often the design of such studies follows particular patterns that have become traditional. Performing Monte Carlo studies is usually not taught as such in graduate schools. As a novice one is simply expected to imitate and extend relevant earlier studies published in the recent literature. Many scholars seem to think that setting up a Monte Carlo study is basically too self-evident to bother much about; apparently, it can be done without requiring a manual, because that does not seem available. Therefore, we try to present and illustrate the fundamentals of executing such studies here, pointing to opportunities not often utilized in current practice, especially regarding designing their general setup, controlling their accuracy, recognizing their shortcomings, presenting their results in a coherent and palatable way, and with respect to an appropriate and unprejudiced interpretation of their actual findings.

Monte Carlo simulation (abbreviated as MCS from now on) produces from random experiments rather straight-forward statistical inference on the properties of often very complex statistical inference techniques. So, it has an intrinsic recursive nature, because it employs

statistical methods to explore statistical methods. Here, we will focus in particular on exploring the properties of classic econometric inference techniques by simulation. The major issues concerning these techniques are concisely characterized in Appendix A. In practice, they are usually applied to observational (i.e., non-experimental) data, employing methods and making probabilistic assumptions in a situation of high uncertainty regarding the appropriate model specification. Hence, in this context MCS examines complex techniques of statistics by rather simple techniques of statistics, aiming to produce knowledge on how to handle non-experimental data by experimentation.

At first sight such an approach may seem to be built on very weak methodological grounds, if not just being impossible. Indeed, its inherent circularity and apparent incompatibilities may easily lead to confusion. Reasons for that being that concepts such as sample and its sample size, estimators and their precision, test statistics and their significance levels, confidence regions and their coverage probabilities, and so on, play a role at two different levels, namely that of the econometric technique under investigation and that of the simulation inference produced on its properties. Therefore, we shall find it useful to develop a notation in which we carefully distinguish between the econometric issues under study and the statistical inference methods employed in MCS to interpret the simulation experiments. Such a distinction is nonstandard in the literature, but we think it is illuminating and certainly useful from a pedagogic point of view. For similar reasons we find it also highly instructive to use EViews programs for illustrations. Not because we appreciate the EViews programming language as such very much, but because it will prove to be most clarifying and convenient that in the computer sessions to follow we will as a rule have two data workfiles. One regarding samples to which the econometric techniques under study are applied, and one with usually a much larger sample size concerning the executed simulation experiments. To the latter we may immediately (or at a later stage, and step by step) apply any descriptive or inferential statistical techniques from the standard EViews menu deemed useful for interpreting the simulation findings.

In the first three sections the focus is on the basic tools of MCS, which are generating and transforming random numbers such as may

arise in econometric analysis, and next assessing their moments, probability distributions and their quantiles numerically. We discuss and illustrate their use to produce MCS inference on the qualities of various specific econometric inference techniques, and to control the accuracy of the MCS results. Especially regarding the accuracy of MCS results we produce some findings that are seldom employed in practice. But also regarding some more familiar results we think that by carefully distinguishing in our notation between the statistical inference techniques under study and the statistical inference techniques employed to interpret the simulation experiments, we illuminate various aspects that are easily confused or overlooked otherwise. At the same time, by illustrating MCS to various of the standard tools of econometric inference, one may acquire a deeper and more tangible understanding of often rather abstract aspects of econometric theory. Not only does it illustrate the relevance and accuracy (and often the inaccuracy, thus limited relevance) of asymptotic theory and of the approximations it suggests. It will also help to sharpen the understanding of basic concepts such as bias and inconsistency, standard deviation and standard error, variance and mean squared error, the standard deviation of standard errors, nominal and actual significance levels, test size and power, and to appreciate how crucial (or occasionally trifling) the validity of particular standard assumptions (such as exogeneity, linearity, normality, independence, homoskedasticity) may be. Both producers and consumers of MCS results may appreciate the easy-to-use rules of thumb provided on the relationship between accuracy of the various obtained MCS results such as bias, median, RMSE, rejection probability, and the chosen number of replications or the Monte Carlo sample size.

After treating the basic tools of MCS, the focus of Section 4 is on the crucial elements of analyzing the properties of asymptotic test procedures by MCS. This involves verifying the extent of control over the type I error probability, establishing the test size, essential aspects of size correction when making power comparisons between competing test procedures. In Section 5 the focus is on various more general aspects of MCS, such as its history, possibilities to increase its efficiency and effectivity, whether synthetic random exogenous variables should be kept fixed over all the experiments or be treated as genuinely random

and thus redrawn every replication. Here we also pay some attention to what we call a dedicated MCS study. Finally, it tries to provide a list of all methodological aspects that do affect MCS. We pay attention especially to those which are relevant when simulation results are used to rate various alternative econometric techniques. Most of these aspects receive very little attention in the majority of the currently published simulation studies. We list ten general methodological rules and aspirations, or rather commandments, to be followed when designing and executing Monte Carlo studies in econometrics when its purpose is an impartial validation of alternative inference techniques. Next we address the adverse effects sinning against these rules has.

The simulation techniques that we discuss in the first five sections are often addressed as naive or classic Monte Carlo methods. However, simulation can also be used not just for assessing the qualities of inference techniques, but also directly for obtaining inference in practice from empirical data. Various advanced inference techniques have been developed which incorporate simulation techniques. An early example of this is Monte Carlo testing, which corresponds to the (much later developed) parametric bootstrap technique. In the final Section 6 such techniques are highlighted, and a few examples of (semi-)parametric bootstrap techniques are given. This section also demonstrates that the bootstrap is not an alternative to MCS but just another practical — though usually asymptotic, and therefore probably inaccurate — inference technique, which uses simulation to produce econometric inference. If one wants to analyze the actual performance of bootstrap inference this can again be done by MCS, as we illustrate. Other advanced uses of simulation, such as in indirect inference or estimation by simulated moments methods or MCMC (Markov chain Monte Carlo) methods will not be covered here.

At the end of each section exercises are provided which allow the reader to immerse in performing and interpreting MCS studies. The material has been used extensively in courses for undergaduate and graduate students. The various sections contain illustrations which throw light on what uses can be made from MCS to discover the finite sample properties of a broad range of alternative econometric methods with a focus on the rather basic models and techniques. Just

occasionally, we pay attention (by giving references) to how to condense the often rather extensive reporting on simulation findings by employing graphical 2D and 3D methods, which can even be extended to 4D by using animation. This, however, requires other software than provided by the EViews package.

Although Monte Carlo is practiced now for more than a century and started in fact long before computers were available by manually drawing repeatedly independent samples from a given population, there are no many texts that explain and thoroughly illustrate MCS and its foundations in detail for econometricians. In that respect we should however name at least the following relatively few exceptions. The relevant underlying theory for examining isolated inference techniques (estimators and test procedures) can be found in Hendry (1984, Section 16 of *Handbook of Econometrics*, Vol. II). Sections on Monte Carlo simulation can also be found in the econometrics textbooks by Davidson and MacKinnon (1993, Section 21), Hendry (1995, Section 3, Section 6), Hendry and Nielsen (2007, Section 16), and intermingled with bootstrap applications throughout most sections of Davidson and MacKinnon (2004). An initial study in Monte Carlo methodology, focussing on issues that are relevant when alternative inference methods are compared by Monte Carlo methods, is Kiviet (2007), which is extended here. A recent introduction to the basics of Monte Carlo methods, focussing in particular on random number generation, is Doornik (2006), published in a volume in which Davidson and MacKinnon (2006) provide an introduction for econometricians to the bootstrap. Further relatively recent introductions to the bootstrap are Horowitz (2003), Johnson (2001), and MacKinnon (2002, 2006). There are many more advanced bootstrap papers in the econometrics literature, see for instance Brown and Newey (2002). For other inference methods which involve simulation (i.e., where simulation is not just used to analyze the quality of inference but to help to produce inference), which are not covered here, such as Method of Simulated Moments, Indirect Inference, Markov Chain Monte Carlo, Gibbs Sampling, Simulated Annealing etc. see, for instance, Fishman (2006) and Gourieroux and Monfort (1996) and the overview in Greene (2012).

Acronyms and Symbols Used

AR(p)	autoregressive process of order p
ARX(p)	regression model with exogenous regressors X and lagged dependent variables up to order p
BSS	bootstrap simulation
CDF	cumulative distribution function
CLT	central limit theorem
DGP	data generating process
E	expectation
ECDF	empirical cumulative distribution function
EPDF	empirical probability distribution function
ϵ	relative precision
IQR	interquartile range
GLS	generalized least-squares
GMM	generalized method of moments
IID	identically and independently distributed
IV	instrumental variables
κ	kurtosis
LIE	law of iterated expectations

LLN	law of large numbers
λ	skewness
MCS	Monte Carlo simulation
MD	median
ML	maximum likelihood
MSE	mean squared error
NIID	normal and IID
NLS	nonlinear least-squares
OLS	ordinary least-squares
PDF	probability density function
PMC	pivotal Monte Carlo (test)
RMSE	root mean squared error
SD	standard deviation
SE	standard error
τ	absolute tolerance
TSLS	two-stage least-squares
UIID	uniform and IID
Var	variance
VSE	variance of the squared error

1

Introduction to Classic Monte Carlo Simulation

1.1 Main Purposes and Means

Computers have the facility to generate seemingly independent drawings from well-known discrete and continuous distributions, such as Bernoulli, binomial, uniform, normal, Student, etc. using a so-called pseudo random number generator. By transforming these random numbers it is also possible to obtain drawings from the distribution of complicated functions of such standard distributions, and thus for any econometric estimator of a parameter vector, and for its variance estimator and for related test statistics. The analytic assessment of the actual cumulative distribution functions, densities, quantiles or moments of these estimators and test statistics is usually intractable, because mostly they are highly nonlinear functions of the random disturbances, the often random regressor and instrumental variables and the model parameters. By using the computer to draw a large IID (identically and independently distributed) sample from such a complicated distribution, we can use this Monte Carlo sample to estimate its moments numerically, provided these exist, whereas a histogram of this sample establishes the empirical probability distribution. Likewise, the empirical counterparts of the cumulative distribution

function (CDF), and if this exists the probability density function (PDF) can be assessed, and quantiles of the unknown distribution can be found by inverting the CDF. Of course, such Monte Carlo estimators of characteristics of an unknown distribution do entail estimation errors themselves. So, in order to be able to judge their inaccuracies, Monte Carlo results should — like all statistical inference — always be supplemented by appropriate corresponding standard errors, confidence regions, etc.

In this introductory section, we will not yet practice Monte Carlo simulation properly, but just illustrate the generation of random variables on a computer by EViews, and employ this to illustrate various basic aspects of the LLN (Law of Large Numbers) and the CLT (Central Limit Theorem), which jointly do not only form the backbone of asymptotic theory on econometric inference, but — as we shall soon find out — also of the interpretation and the control of the actual accuracy of MCS (Monte Carlo simulation) findings.

1.2 Generating Pseudo Random Numbers

A digital computer cannot really generate genuinely random numbers nor throw dices. Though, it can generate series of so-called pseudo random numbers in the (0, 1) interval by applying a deterministic algorithm to an initial positive integer number called the seed. If one knows this seed and the algorithm all drawings are perfectly predictable, but if not, they have the appearance of IID drawings from the uniform distribution over the 0-1 interval. If one just knows the seed but not the algorithm one cannot predict the series, but by using the algorithm with the same seed again one can reproduce the same pseudo random series whenever desired. This comes in handy when one wants to compare alternative methods under equal circumstances.

The algorithm for producing random numbers is usually of the following simple type. Let $z_0 > 0$ be the positive integer seed, then pseudo IID $U(0,1)$ drawings $(\eta_1, \eta_2, ...)$ follow from the iterative scheme

$$\left.\begin{array}{c} z_i = (\psi z_{i-1} + \alpha) \div m \\ \eta_i = z_i/m, \end{array}\right\} \quad i = 1, 2, \ldots \tag{1.1}$$

where integer m is called the modulus, ψ is the multiplier, and α the increment. The operation $\div$ (often denoted as *mod*) means here that z_i equals the remainder of dividing $\psi z_{i-1} + \alpha$ by m. In at most m steps the series of values z_i and thus η_i will repeat itself. Hence, m should be large, preferably as large as the largest integer on the computer, say $2^{31} - 1$. The choice of the value of ψ is crucial for the quality of the series too, but α is less relevant and is often set at zero. Testing the adequacy of random number generators is an art of its own. We will simply trust the default versions available in the computer package that we use.

The CDF of $\eta_i \sim U(0,1)$ is

$$F_U(\eta) \equiv \Pr(\eta_i \leq \eta) = \begin{cases} 0, & \eta < 0 \\ \eta, & 0 \leq \eta \leq 1 \\ 1, & \eta > 1. \end{cases} \tag{1.2}$$

By appropriately transforming IID drawings $\eta_i \sim U(0,1)$ one can obtain IID drawings ζ_i from any other type of distribution D with strictly increasing CDF given by $F_D(\zeta)$, with $\zeta \in \mathbb{R}$. Consider $\zeta_i = F_D^{-1}(\eta_i)$. This yields $F_D(\zeta_i) = \eta_i \sim U(0,1)$. So

$$\Pr(\zeta_i \leq \zeta) = \Pr(F_D(\zeta_i) \leq F_D(\zeta)) = \Pr(\eta_i \leq F_D(\zeta)) = F_D(\zeta) \tag{1.3}$$

indeed. Hence, generating $\zeta_i = F_D^{-1}(\eta_i)$ yields a series of IID pseudo random drawings of ζ_i, $i = 1, 2, \ldots$. This does not work out well when distribution D has a CDF that has no closed form for its inverse, as is the case for the (standard) Normal distribution. However, relatively simple alternative transformation techniques are available for that situation, see for instance Fishman (2006).

1.2.1 Drawings from $U(0,1)$

We shall now use EViews[1] to illustrate the above and interpret some realizations of series of generated random drawings. At the same time

[1] Check the EViews reference guide to find out about any particulars on the random number generator that your version of EViews uses. All results to follow were obtained by EViews 7. Earlier versions use a different random number generator and therefore do not yield results fully similar to those presented here.

we shall learn how to use the programming facilities of EViews. Consider the following EViews program:

```
'mcs11.prg:  Drawings from U(0,1)
!n=1000
workfile f:\MCS\mcs11.wf1 u 1 !n
rndseed 9876543210
genr eta=rnd
```

The first line of this program[2] (and all programs to follow) starts with the character "'" meaning that it just contains a comment and (without any computational consequences) exhibits the name (with extension prg) and purpose of the program. Names of integer or real variables have to be preceded by the character "!". By !n we identify sample size. In the third line we identify a new workfile and its location (map or folder); for clarity we give this workfile (which has extension wf1) the same name as the program. The parameter "u" indicates that the observations are "undated" (not associated with a particular year or quarter) and next it is indicated that their range will run form 1 to !n. In the fourth line we provide a specific but arbitrary integer seed value for the random number generator, and in the final line a variable eta is generated of n IID drawings from $U(0,1)$. After running this program in an EViews session one can manipulate and analyze the data series eta stored in the workfile as one wishes, either by using standard EViews commands or by running another program operating on this workfile.

Figure 1.1 presents the histograms, as produced by EViews, obtained from running program mcs11 first for $n = 1{,}000$ and next for $n = 1{,}000{,}000$. Both these histograms establish empirical probability density functions of the $U(0,1)$ distribution. Both deviate from the rectangular actual population PDF, and obviously and visibly the one with larger n is more accurate. The value of mean is calculated according to

$$\bar{\eta}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\eta_i. \tag{1.4}$$

[2] All programs are available in a zipped file mcs.zip. These programmes suppose that you have access to a drive f:\ with folder MCS. Of course, the path f:\MCS\ can be changed in whatever is more convenient.

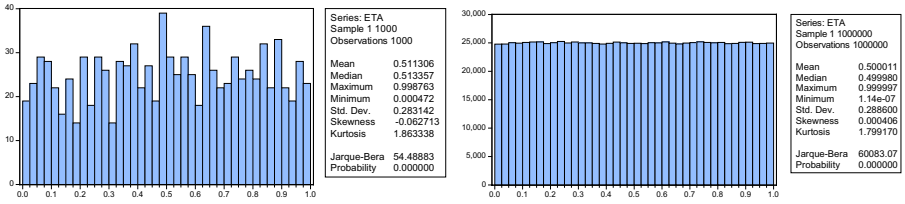

Fig. 1.1 Drawings from $U(0,1)$ for $n = 1{,}000$ and $n = 1{,}000{,}000$.

For both sample sizes these are pretty close to $E(\eta) = 0.5$. The reason is that, when the η_i establish a series of $U(0,1)$ drawings indeed, the sample average $\bar{\eta}_n$ is an unbiased estimator of $E(\eta)$, because

$$E(\bar{\eta}_n) = E\frac{1}{n}\sum_{i=1}^{n}\eta_i = \frac{1}{n}\sum_{i=1}^{n}E(\eta_i) = \frac{n}{n}E(\eta) = 0.5. \tag{1.5}$$

Of course, the actual deviation of mean from 0.5 is associated with its standard deviation. We find

$$\begin{aligned}\mathrm{Var}(\bar{\eta}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\eta_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n}\eta_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(\eta_i) = \frac{1}{n}\mathrm{Var}(\eta) = \frac{1}{12n},\end{aligned} \tag{1.6}$$

where the third equality follows from the independence of the drawings, yielding $\mathrm{Cov}(\eta_i, \eta_j) = 0$ for $i \neq j$, and the final one from using general results on the $U(a,b)$ distribution. Regarding its first four centered moments we have $\mu_1^c \equiv E[\eta - E(\eta)] = 0$, $\mu_2^c \equiv \sigma_\eta^2 \equiv E[\eta - E(\eta)]^2 = (b-a)^2/12$, $\mu_3^c \equiv E[\eta - E(\eta)]^3 = 0$, and $\mu_4^c \equiv E[\eta - E(\eta)]^4 = (b-a)^4/80$, respectively. Since $\sqrt{\mathrm{Var}(\bar{\eta}_n)} = (12n)^{-1/2} = 0.0091$ and 0.00029, for $n = 1{,}000$ and $n = 1{,}000{,}000$ respectively, we can understand why the Mean value is much closer to 0.5 in the larger sample. Note, however, that there will be values of the seed for which the sample mean for $n = 1{,}000$ is closer to 0.5, because it is not the deviation itself that will be larger at $n = 1{,}000$ than at $n = 1{,}000{,}000$, but for a smaller sample size the probability is larger that the deviation will be larger than a particular value. For similar reasons it

is understandable that the estimate of the median is more accurate for the larger n.

The value of Std. Dev. (which is actually what we would usually call the standard error, because it is the estimated standard deviation) mentioned next to the histograms is obtained as the square root of

$$\hat{\sigma}_\eta^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta}_n)^2. \tag{1.7}$$

Both estimates are pretty close to the standard deviation of the $U(0,1)$ distribution, which is $\sqrt{1/12} = 0.2887$. This is again no surprise, because

$$E(\hat{\sigma}_\eta^2) = \sigma_\eta^2 = \mathrm{Var}(\eta) = 1/12. \tag{1.8}$$

Note that the definitions of skewness and kurtosis are $\mu_3^c/(\mu_2^c)^{3/2}$ and $\mu_4^c/(\mu_2^c)^2$ respectively, so the population values for the $U(0,1)$ distribution are 0 and $144/80 = 1.8000$, respectively. Again we find that the estimates obtained from the two samples are pretty close to their population values, and they are closer for the larger sample size. The improvements with n are due to the fact that the corresponding estimators do have a variance of order $O(n^{-1})$.

1.2.2 Drawings from $N(0,1)$

Next we adapt the program as follows:

```
'mcs12.prg  Drawings from N(0,1)
!n=1000
workfile f:\MCS\mcs12.wf1 u 1 !n
rndseed 9876543210
genr zeta=nrnd
```

This yields for $n = 1{,}000, 1{,}000{,}000$ the histograms of Figure 1.2. The results on these samples of $N(0,1)$ drawings can be analyzed in a similar way as we did for $U(0,1)$.[3] Here, however, we have $\mu_1 = 0$, $\mu_2 = 1, \mu_3 = 0$, and $\mu_4 = 3$.

[3] Note that $(0,1)$ indicates the domain for the uniform distribution, whereas it refers to expectation and variance in case of the normal.

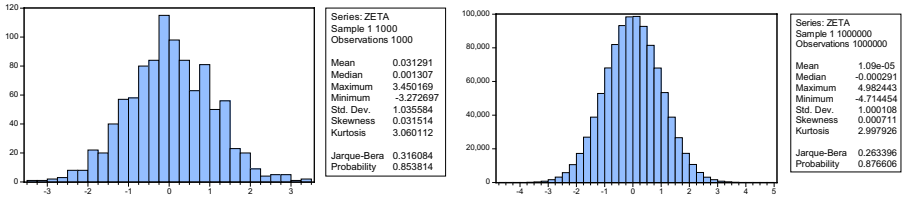

Fig. 1.2 Drawings from $N(0,1)$ for $n = 1{,}000$ and $n = 1{,}000{,}000$.

Note that the phenomenon that the values of mean converge to zero for increasing values of n illustrate the simplest form of the LLN (law of large numbers). Since both η_i and ζ_i are IID and have finite moments the sample average converges to the population mean (expectation). The same holds for the uncentered sample averages of powers of η_i and ζ_i which explains, upon invoking Slutsky's theorem, the convergence of their nonlinear transformations Std. Dev., Skewness and Kurtosis. See Appendix B for more details on the various tools (the notation $O(n^{-1})$, LLN, CLT, Slutsky) for asymptotic analysis.

1.3 LLN and Classic Simple Regression

Both in MCS and in econometrics the LLN plays a central role. Therefore, to understand its properties better, we will now provide some illustrations on the workings of the LLN in the context of a very simple regression model (and also on its limitations in Exercise 9). We consider the model with just one single exogenous regressor and start with the case where the observations are IID, and not necessarily normal. For $i = 1, \ldots, n$ the DGP (data generating process) is

$$y_i = \beta x_i + u_i, \quad x_i \sim IID(\mu_x, \sigma_x^2), \quad u_i \mid x_1, \ldots, x_n \sim IID(0, \sigma_u^2), \tag{1.9}$$

where μ_x, σ_x^2, and σ_u^2 are all finite. A well-known asymptotic (for $n \to \infty$) result in econometric theory is that in this model, due to the LLN, $\operatorname{plim} n^{-1} \sum_{i=1}^n x_i^2 = \lim n^{-1} \sum_{i=1}^n E(x_i^2) = \sigma_x^2 + \mu_x^2$ and $\operatorname{plim} n^{-1} \sum_{i=1}^n x_i u_i = \lim n^{-1} \sum_{i=1}^n E(x_i u_i) = 0$, since by the LIE (Law of Iterated Expectations) $E(x_i u_i) = E[E(x_i u_i \mid x_i)] = E[x_i E(u_i \mid x_i)] = E(0) = 0$. Employing Slutsky, we now find consistency for the OLS

estimator, because

$$\begin{aligned}\operatorname{plim}\hat{\beta} &= \operatorname{plim}\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\ &= \beta + \frac{\operatorname{plim} n^{-1}\sum_{i=1}^{n} x_i u_i}{\operatorname{plim} n^{-1}\sum_{i=1}^{n} x_i^2} = \beta + \frac{0}{\sigma_x^2 + \mu_x^2} = \beta. \qquad (1.10)\end{aligned}$$

Of course, on a computer we cannot fully mimic the situation $n \to \infty$, but the analytic phenomena just discussed are nevertheless convincingly (especially when you increase the value of nmax) illustrated by the following program.

```
'mcs13.prg: LLN in simple IID regression
!nmax=1000
workfile f:\MCS\mcs13.wf1 u 1 !nmax
!beta=0.5
!mux=1
!sigx=2
!sigu=0.2
rndseed 9876543210
genr x=!mux + !sigx*(rnd - 0.5)/@sqrt(1/12)
genr u=!sigu*(rnd - 0.5)/@sqrt(1/12)
genr y=!beta*x + u
stom(x,vecx)
stom(y,vecy)
matrix (!nmax,3) results
!sumxy=0
!sumxx=0
for !n=1 to !nmax
  !sumxy=!sumxy+vecx(!n)*vecy(!n)
  !sumxx=!sumxx+vecx(!n)^2
  results(!n,1)=!sumxy/!n
  results(!n,2)=!sumxx/!n
  results(!n,3)=!sumxy/!sumxx
next
results.write f:\MCS\mcs13results.txt
read f:\MCS\mcs13results.txt sxx sxy b
```

In this program we chose $\beta = 0.5$, $x_i \sim IID(1, 2^2)$, $u_i \sim IID(0, 0.2^2)$ and both x_i and u_i are uniformly distributed and mutually independent. By the command "stom" we transform a generated series into a vector, which then enables to program expressions involving its individual elements. This allows to calculate both $n^{-1}\sum_{i=1}^{n} x_i^2$ and $n^{-1}\sum_{i=1}^{n} x_i y_i$ for $n = 1, \ldots, 1000$, and also their ratio. The results are stored in a matrix called results, which has nmax rows and 3 columns. In the two final lines of the program this 1,000 × 3 matrix is first written to a text file and from this the three columns are added as variables to the workfile mcs13.wf1 under the names sxy, sxx and b. These can then easily be analyzed further by EViews.

Figure 1.3 presents the graphs of sxy, sxx, and b as obtained for two different integer seed values. For small values of n the random nature of the three depicted statistics is apparent from the diagram, but they gradually converge for increasing n (see Exercise 6), and ultimately for $n \to \infty$, irrespective of the value of rndseed used, they assume their deterministic population values which are $\sigma_x^2 + \mu_x^2 = 5$, $\text{plim}\, n^{-1}\sum_{i=1}^{n} x_i y_i = \beta \text{plim}\, n^{-1}\sum_{i=1}^{n} x_i^2 = 2.5$, and $\beta = 0.5$ respectively. In fact, due to the correlation between sxy and sxx we note that the convergence of their ratio is much faster than that of sxy and sxx individually.

The LLN does not require that the observations in the regression are IID; they only have to be asymptotically uncorrelated. We will now verify the effects of first-order serial correlation in both the

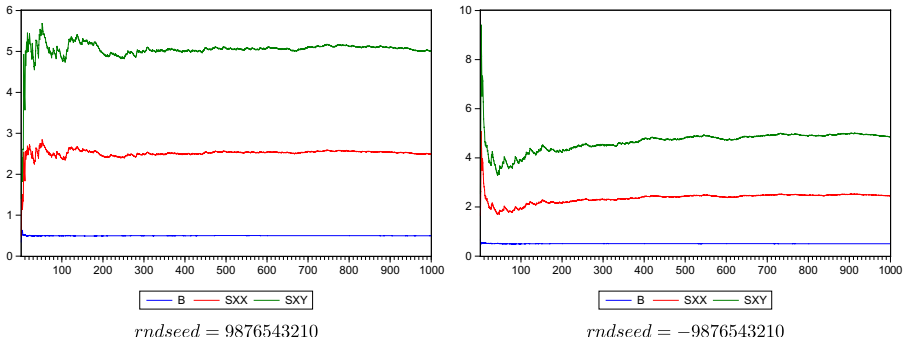

Fig. 1.3 IID data; second moments (uncentered) and $\hat{\beta}$ for $n = 1, \ldots, 1{,}000$.

regressor x_i and the disturbances u_i. Because EViews generates variables recursively, in program msc14.prg variables $x_i^* \sim IID(0, \sigma_x^2)$ and $u_i^* \sim IID(0, \sigma_u^2)$ are actually generated first (but already stored in x_i and u_i) for $i = 1, \ldots, n\max$ and then the program calculates $u_1 = u_1^*$ and (for $i > 1$) $u_i = \rho_u u_{i-1} + (1 - \rho_u^2)^{1/2} u_i^*$ and similarly for x_i, to which finally μ_x is added.

```
'mcs14.prg: LLN in simple non-IID regression
!nmax=1000
workfile f:\MCS\mcs14.wf1 u 1 !nmax
!beta=0.5
!mux=1
!sigx=2
!rhox=0.8
!rrhox=@sqrt(1-!rhox^2)
!sigu=0.2
!rhou=0.4
!rrhou=@sqrt(1-!rhou^2)
rndseed 9876543210
genr x=!sigx*(rnd - 0.5)/@sqrt(1/12)
genr u=!sigu*(rnd - 0.5)/@sqrt(1/12)
smpl 2 !nmax
genr x=!rhox*x(-1)+!rrhox*x
genr u=!rhou*u(-1)+!rrhou*u
smpl 1 !nmax
genr x=!mux + x
genr y=!beta*x + u
stom(x,vecx)
stom(y,vecy)
matrix (!nmax,3) results
!sumxy=0
!sumxx=0
for !n=1 to !nmax
  !sumxy=!sumxy+vecx(!n)*vecy(!n)
  !sumxx=!sumxx+vecx(!n)^2
  results(!n,1)=!sumxy/!n
```

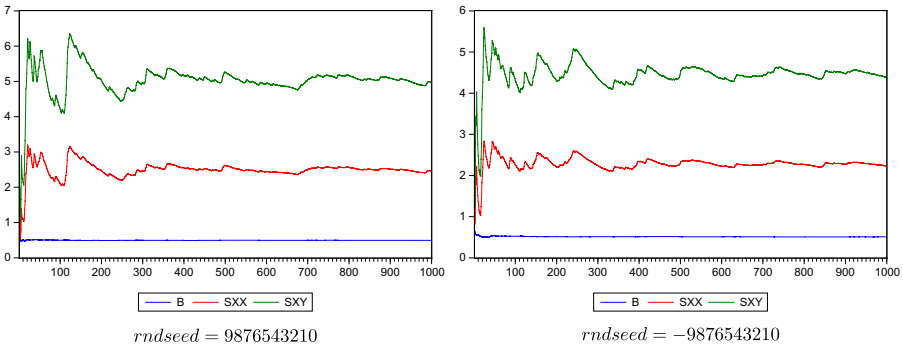

Fig. 1.4 Non-IID data; second moments (uncentered) and $\hat{\beta}$ for $n = 1,\ldots,1{,}000$.

```
  results(!n,2)=!sumxx/!n
  results(!n,3)=!sumxy/!sumxx
next
results.write f:\MCS\mcs14results.txt
read f:\MCS\mcs14results.txt sxx sxy b
```

Figure 1.4 illustrates that OLS is also consistent in a model where either the observations of the regressors or those of the disturbances (or both) are serially correlated, but more and more uncorrelated at greater distance, provided that regressors and disturbances are contemporaneously uncorrelated. One can derive

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}x_iu_i\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}x_iu_i\right)^2 = \frac{1}{n^2}E\left(\sum_{i=1}^{n}x_iu_i\right)^2 \\ &= \frac{\sigma_x^2\sigma_u^2}{n}\left(\frac{1+\rho_x\rho_u}{1-\rho_x\rho_u} - 2\frac{\rho_x\rho_u}{n}\frac{1-(\rho_x\rho_u)^n}{(1-\rho_x\rho_u)^2}\right),\end{aligned}$$

from which it follows that the convergence for the numerator, although of the same rate in n, is nevertheless slower for positive ρ_x and ρ_u than in the IID case, where $\rho_x = \rho_u = 0$. This is clearly illustrated by the diagrams.

In both programs mcs13.prg and mcs14.prg the regressor variable is strongly exogenous, because $E(u_i \mid x_1,\ldots,x_n) = 0 \ \forall i$, and the disturbances are homoskedastic. Note that in establishing the consistency of OLS we only used $E(u_i \mid x_i) = 0$. So, neither serial correlation nor

heteroskedasticity of the disturbances would spoil this result. Also weak exogeneity or predeterminedness of the regressors, where $E(u_i \mid x_1, \ldots, x_i) = 0$, still yields consistency of OLS.

1.4 CLT and Simple Sample Averages

The conclusions that we will draw on the basis of MCS studies will rely heavily, as far as their accuracy is concerned, on a very straight-forward application of the simplest version of the Central Limit Theorem. In MCS we will often approximate the sample average of IID observations generated from a usually unknown distribution by the normal distribution. In fact, we will standardize the sample average and approximate the outcome with the standard normal distribution. This approximation is perfect when it concerns a sample of NIID observations or when the sample is infinitely large, but it will involve approximation errors when the sample observations have a nonnormal distribution and the sample size is finite. In the illustration to follow we will examine the quality of the approximation for a few different nonnormal distributions and for various finite sample sizes.

Program mcs15.prg calculates sample averages from a sample of size n for five different IID variables. These variables are ($i = 1, \ldots, n$): $z_i \sim N(0,1)$, $v_i \sim N(\mu_v, \sigma_v^2)$, $u_i \sim U(a,b)$, $x_i \sim \chi^2(2)$ and w_i, where the latter is a mixture of independent $\chi^2(1)$ and $\chi^2(2)$ variables. The generation of samples of size n of these variables, and the calculation of the sample averages is replicated R times. Running the program results in two workfiles. Every replication workfile mcs15.wf1 contains the n observations on the five variables, and after termination of the program these are their realizations in the final replication. Workfile mcs15res.wf1 contains variables of R observations. These are the generated sample averages and also their rescaled versions, in deviation from their expectation and divided by the standard deviation of the sample average. Of course, the latter equals the standard deviation of the individual elements divided by $\sqrt{n}$ (prove this yourself!). The CLT implies that for $n \to \infty$ the rescaled expressions should be indistinguishable from drawings from the $N(0,1)$ distribution. By the program we will examine how close we get to that when n is 10, 100, or 1,000.

In fact, we run the program first for $n = 1$, not because we expect the CLT to have much to say then, but because this is a straight-forward way to obtain the histograms of R drawings from the distributions of the five different random variables from this program.

```
'mcs15.prg:  Sample averages and the CLT
!n=10
workfile f:\MCS\mcs15.wf1 u 1 !n
!muv=1
!sigv=3
!a=5
!b=15
rndseed 9876543210
!R=10000
matrix (!R,5) simres
for !rep=1 to !R
   genr z=nrnd
   genr v=!muv + !sigv*z
   genr u=!a + (!b-!a)*rnd
   genr x=@rchisq(2)
   genr p=rnd>0.75
   genr w=p*(x+3) - (1-p)*(z^2+2)
   simres(!rep,1)=@mean(z)
   simres(!rep,2)=@mean(v)
   simres(!rep,3)=@mean(u)
   simres(!rep,4)=@mean(x)
   simres(!rep,5)=@mean(w)
next
simres.write f:\MCS\mcs15res.txt
workfile f:\MCS\mcs15res.wf1 u 1 !R
read f:\MCS\mcs15res.txt meanz meanv meanu meanx meanw
genr rescaledmeanz=@sqrt(!n)* meanz
genr rescaledmeanv=@sqrt(!n)*(meanv - !muv)/!sigv
genr rescaledmeanu=@sqrt(!n)*(meanu - (!a + !b)/2)/((!b - !a)/@sqrt(12))
genr rescaledmeanx=@sqrt(!n)*(meanx - 2)/2
genr rescaledmeanw=@sqrt(!n)*(meanw + 1)/@sqrt(14.5)
```

In Figure 1.5 histograms are shown of 10,000 drawings from $v \sim N(1, 3^2)$, $u \sim U(5, 15)$, $x \sim \chi^2(2)$ and w, where the latter two are non-symmetric and the last one is clearly bimodal. Note that u has thin tails (low kurtosis), that x is skew to the right and has very high kurtosis, whereas w also has positive skewness and kurtosis larger than for the normal distribution. The Jarque–Bera test notes indeed that the latter three distributions are seriously nonnormal.

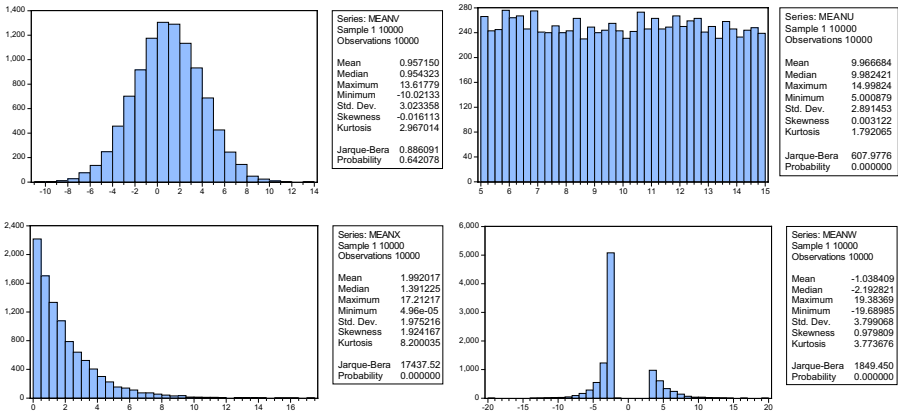

Fig. 1.5 Histograms of 10,000 drawings of v, u, x, and w.

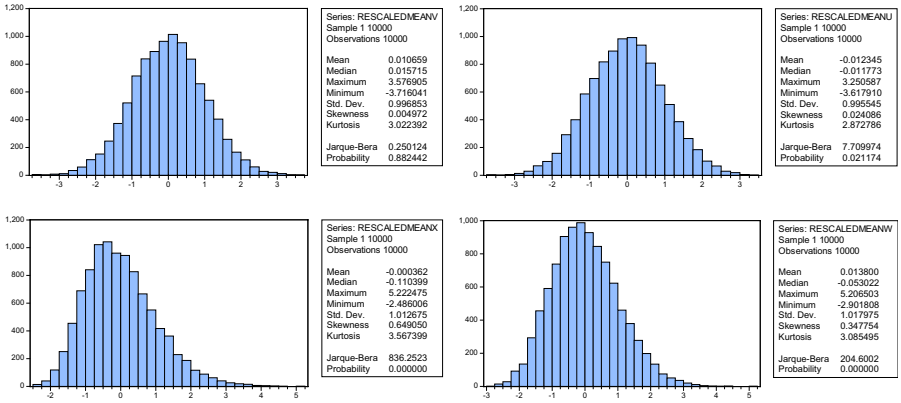

Fig. 1.6 Distribution of rescaled sample averages for $n = 10$.

In Figure 1.6, $n = 10$ and histograms of the rescaled sample averages are presented. From the histograms of 10,000 drawings it is obvious that even at $n = 10$ these sample averages already started to converge toward the normal. This is self-evident for rescaledmeanv, because these are drawings from the standard normal distribution for any n. The symmetry of the uniform distribution implies that rescaledmeanu has skewness very close to zero for any n, and at $n = 10$ the kurtosis is already such that the normality hypothesis, although actually invalid, is not rejected at the 1% level. Also for the sample averages of the x

and w distributions the skewness and the kurtosis are already much closer to zero and three respectively, although still such that normality is strongly rejected by the Jarque–Bera test.

Figure 1.7 contains results for $n = 100$ and shows that averaging has completely removed the underlying nature of the uniform drawings and of the bimodality of the w_i drawings, but the skew nature of the x_i and w_i distributions is still emerging from the averages of samples of this size.

Continuing this and taking $n = 1{,}000$ it is shown in Figure 1.8 that the sample averages are again closer to normal. These illustrations

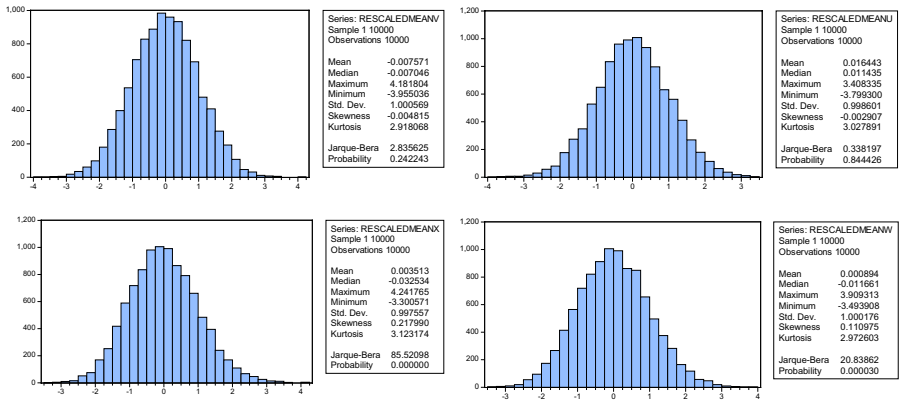

Fig. 1.7 Distribution of rescaled sample averages for $n = 100$.

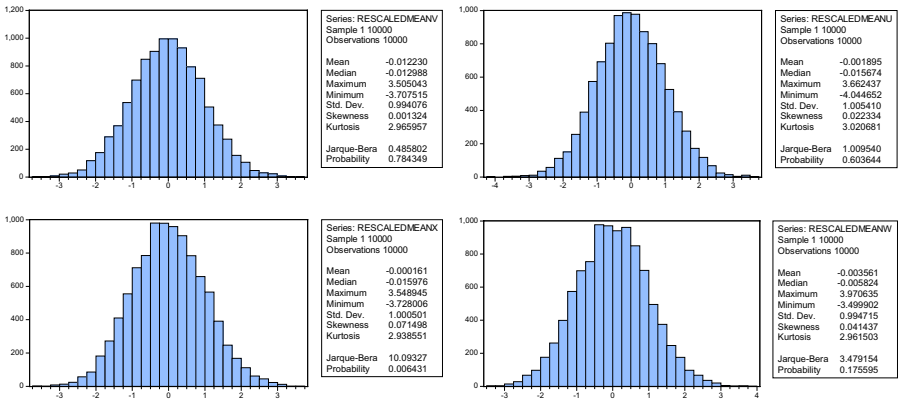

Fig. 1.8 Distribution of rescaled sample averages for $n = 1{,}000$.

clearly demonstrate that both the degree of nonnormality of the underlying random variables and the size of the sample over which the average is taken jointly determine the accuracy of the normal approximation. Seriously nonnormal distributions are shown to have sample averages that are distributed closely to (although possibly still significantly different from) normal when the sample is as large as 1,000. In fact, when it comes to the accuracy of the MCS inferences to be developed in the following sections, it will be argued that it does not matter that much whether or not the distribution of a sample average as such is very accurately approximated by the normal distribution, but only whether its tail probabilities and thus its quantiles in the tail areas conform closely to those of the normal. We will examine that in more detail later, and also other aspects that determine the accuracy of MCS inference.

In most of the simulated above results, we obtained information on relatively simple statistics for which many of their typical properties, especially their moments, can be derived analytically. So, the simulation results merely serve as a specific numerical illustration of already fully understood more general characteristics. This will also be the case in the more involved MCS illustrations on particular parametrizations of the standard normal linear regression model in the next section. The distribution of their relevant statistics can be derived analytically, so there is no genuine need for simulating them other than illustration. However, such results help to appreciate that MCS results lack generality, are nonanalytical but numerical, and are random as well. Therefore, they are both very specific and involve inaccuracies, and only after fully understanding the nature and magnitude of these inaccuracies we will move on and apply MCS in a range of situations where the true underlying properties of estimators and tests are mostly unknown and the estimated numerical results from MCS establish, next to their analytic asymptotic approximations, our only guidebook.

Exercises

1. Consider (1.8). Prove $E(\hat{\sigma}_\eta^2) = \text{Var}(\eta)$. Also, from $F_U(\eta)$ in (1.2), find the density of $U(0,1)$ and derive $\text{Var}(\eta) = 1/12$.

2. Explain the high values of the Jarque–Bera statistics (consult the EViews Help facility for more details on this test) in Figure 1.1 and argue why these lead to low p-values. General warning: Never mistake the p-value of a test statistic as expressing the probability that the null hypothesis is true, because it simply expresses the probability (according to its asymptotic null distribution) that the test statistic may assume a value as extreme as it did (hence, supposing that the null hypothesis is true).
3. Consider the estimator $\hat{\sigma}_\eta^2$ of (1.7) where $\eta_i \sim UIID(0,1)$. Derive $\text{Var}(\hat{\sigma}_\eta^2)$ and show that it is $O(n^{-1})$.
4. Explain the moderate values of the Jarque–Bera statistics in Figure 1.2 and their corresponding high p-values.
5. Run similar programs as mcs11.prg and mcs12.prg with different values of seed and similar (and also larger) values of n and explain your findings.
6. Run programs mcs13.prg and mcs14.prg for $n = 100,000$ upon changing rndseed and possibly also choosing different values for the parameters μ_x, σ_x, and σ_u. Compare the sample equivalents of μ_x, σ_x, and σ_u with their population values. Also interpret the correlogram of the u and x series.
7. Run program mcs15.prg for $n = 1{,}000$ to replicate the results of Figure 1.7. Consider the histogram of variable meanw. Derive analytically that $E(w) = -1$ and $\text{Var}(w) = 14.5$. Explain why the histograms of meanw and rescaledmeanw show similar values for skewness and kurtosis. Give command genr rejw=abs(rescaledmeanw)>1.96 in the command window of EViews and examine the histogram of variable rejw. Note that, although the distribution of rescaledmeanw does not correspond to the standard normal in all aspects, the probability that a drawing is larger in absolute value than the 2.5% quantile of the standard normal does not seem to differ much from 5%.
8. In Figure 1.8 normality of rescaledmeanu is not rejected. What will be the effect on this finding from running the program for $n = 1{,}000$ with a much larger value of R? And

the same question for any R but a much larger value of n? Try and explain.

9. Run program mcs16.prg (given below) and choose various different rndseed values while keeping the number of degrees of freedom v fixed and examine the histogram of the generated series of random variables for $v = 3, 2, 1$. Note that for a random variable following the *Student*(v) distribution only the moments up to $v - 1$ exist. Its higher-order moments are defined by an integral which is infinite, because the tails of its density function are relatively fat. So, for $v = 3$ the skewness and kurtosis estimates do not converge (the LLN does not apply!). Irrespective of the value of n they yield different outcomes when rndseed is changed. For $v = 2$ the same happens with the standard deviation estimate, and for $v = 1$ (this is the Cauchy distribution) also the sample average has no deterministic population equivalent. Note, however, that the median does converge to zero for n large, irrespective of the values of v and of rndseed.

```
'mcs16.prg       Drawings from Student(v)
!n=10000
workfile f:\MCS\mcs16.wf1 u 1 !n
rndseed 9876543210
!v=3
genr studentv=@rtdist(!v)
```

2

Monte Carlo Assessment of Moments

Inference on particular characteristics of probability distributions obtained by Monte Carlo experiments is based on the following principle. Suppose we are interested in properties of the real and continuous distribution of a scalar statistic q_n, where q_n is a function of a $p \times 1$ parameter vector θ, and also of some known deterministic variables and of particular random variables with known distribution, which pertain to a sample of size n. Hence, q_n may represent an element of an estimator $\hat{\theta}$ of the vector of parameters θ, or an element of an estimator of the variance matrix of $\hat{\theta}$, but it could also represent some scalar test statistic. We assume that the determining random variables can all be collected in a $m \times 1$ vector, say v. This has a known multivariate distribution from which we can draw mutually independent realizations on a computer. Usually $m = O(n)$. Often we simply have $m = n$, for instance in a regression model with fixed regressors and normal disturbances $v = u$ with $u = \sigma_u \varepsilon$, where $\varepsilon \sim N(0, I_n)$. Or $m = n + 1$ in a first-order dynamic regression model with random start-up value y_0 and n disturbances in u, whereas $m = 2n$ in a static simultaneous model for two endogenous variables $y_i^{(1)}$ and $y_i^{(2)}$, and so on.

By choosing particular numerical values for n and θ that we are interested in and for any further relevant collection of determining

observations on deterministic variables, say D (often a $n \times k$ matrix X, sometimes a $n \times l$ matrix of instrumental variables Z, occasionally also deterministic initial conditions y_0), we thus can generate drawings

$$q_n = q_n(\theta, D, v). \tag{2.1}$$

In a Monte Carlo simulation study we generate a usually rather large sample of size R (the number of replications) from the distribution of $q_n(\theta, D, v)$. We make sure that the R drawings

$$(q_n^{(1)}, \ldots, q_n^{(R)}) \tag{2.2}$$

are both stochastically IID. This is easily realized by generating the R vectors v such that they are mutually independent, and by keeping θ, n and the deterministic variables D all fixed over the R replications. Inferences on any population characteristics of the distribution of q_n can now be obtained from analyzing the sample (2.2). First we will focus on assessing moments of q_n, assuming that these exist.

2.1 MCS Estimation of an Expectation

The Monte Carlo sample average, i.e., the arithmetic mean of (2.2),

$$\bar{q}_{R,n} \equiv \frac{1}{R} \sum_{r=1}^{R} q_n^{(r)} \tag{2.3}$$

is an unbiased Monte Carlo estimator of $E(q_n)$, because

$$E(\bar{q}_{R,n}) = \frac{1}{R} \sum_{r=1}^{R} E(q_n^{(r)}) = \frac{R}{R} E(q_n) = E(q_n). \tag{2.4}$$

when q_n is an element, say the jth and denoted as $\hat{\beta}_j$, of a $k \times 1$ vector $\hat{\beta}$, which estimates a corresponding coefficient vector β, then denoting the Monte Carlo estimator of $E(\hat{\beta}_j)$ by putting a bar on top of $\hat{\beta}_j$, as we did in (2.3) with q_n, is notational inconvenient. To clearly distinguish for $j = 1, \ldots, k$ the econometric estimator of β_j from the sample of size n (decorated often, like here, by a hat $\hat{}$), we shall use an arrow $\rightarrow$ to indicate Monte Carlo estimation of $E(\hat{\beta}_j)$ obtained from the sample of size R. We put the arrow on top of the expectation operator in order to indicate that we are not estimating a parameter of

the econometric model as such, but an expectation, which is of course a population parameter of the employed econometric technique in the context of the model and the DGP under consideration. Like q has subindex n to indicate its sample size, we may give $\overrightarrow{E}_R$ subindex R to indicate the Monte Carlo sample size from which the estimated expectation is obtained. For the MCS estimator of $E(q_n)$ this leads to the definition of the unbiased Monte Carlo expectation estimator

$$\overrightarrow{E}_R(q_n) \equiv \frac{1}{R}\sum_{r=1}^{R} q_n^{(r)}, \text{ for which } E[\overrightarrow{E}_R(q_n)] = E(q_n). \tag{2.5}$$

This Monte Carlo estimator $\overrightarrow{E}_R(q_n)$ has variance

$$\begin{aligned}\mathrm{Var}[\overrightarrow{E}_R(q_n)] &= \mathrm{Var}\left[\frac{1}{R}\sum_{r=1}^{R} q_n^{(r)}\right] = \frac{1}{R^2}\mathrm{Var}\left(\sum_{r=1}^{R} q_n^{(r)}\right) \\ &= \frac{1}{R^2}\sum_{r=1}^{R}\mathrm{Var}(q_n^{(r)}) = \frac{1}{R}\mathrm{Var}(q_n),\end{aligned} \tag{2.6}$$

where the third equality follows from the independence of the drawings in the Monte Carlo sample, and the fourth equality from them having identical distributions. Hence, the standard deviation of the Monte Carlo estimator $\overrightarrow{E}_R(q_n)$ of $E(q_n)$ is equal to the standard deviation of q_n divided by $\sqrt{R}$. Thus, by choosing R very large the Monte Carlo inference on $E(q_n)$ can be made arbitrarily accurate! In fact, applying the Law of Large Numbers (LLN) in the Monte Carlo sample, we find

$$\operatorname*{plim}_{R\to\infty} \overrightarrow{E}_R(q_n) = E(q_n), \tag{2.7}$$

for the chosen fixed and finite n. Hence, the MCS estimator of the finite-n expectation is unbiased and large-R consistent. If we were just interested in an unbiased estimator of $E(q_n)$ only one replication ($R = 1$) would be sufficient. It is to increase its precision that we prefer R to be large; as large as required given the aimed at precision, provided that this is computationally feasible.

The Monte Carlo sample can also be used to estimate the second moments $\mathrm{Var}(q_n)$ and $\mathrm{Var}[\overrightarrow{E}_R(q_n)]$. An unbiased Monte Carlo estimator of $\mathrm{Var}(q_n)$, which should be denoted now as $\overrightarrow{\mathrm{Var}}_R(q_n)$, is given

(for $R > 1$) by

$$\overrightarrow{\mathrm{Var}}_R(q_n) \equiv \frac{1}{R-1}\sum_{r=1}^{R}(q_n^{(r)} - \overrightarrow{E}_R(q_n))^2. \tag{2.8}$$

Due to (2.8), self-evidently an unbiased MCS estimator of $\mathrm{Var}[\overrightarrow{E}_R(q_n)]$ is given by

$$\overrightarrow{\mathrm{Var}}_R(\overrightarrow{E}_R(q_n)) \equiv \overrightarrow{\mathrm{Var}}_R(q_n)/R. \tag{2.9}$$

In addition to the LLN, the IID characteristics of the Monte Carlo sample also allow to invoke the Central Limit Theorem (CLT), giving

$$\sqrt{R}(\overrightarrow{E}_R(q_n) - E(q_n)) \underset{R\to\infty}{\overset{d}{\to}} N(0, \mathrm{Var}(q_n)). \tag{2.10}$$

Under regular circumstances the estimator $\overrightarrow{\mathrm{Var}}_R(q_n)/R$, which is unbiased for $\mathrm{Var}(\overrightarrow{E}_R(q_n))$, is also large-$R$ consistent. Now, according to a result attributed to Cramér (see Appendix B), we obtain the limiting distribution

$$\frac{\overrightarrow{E}_R(q_n) - E(q_n)}{(\overrightarrow{\mathrm{Var}}_R(q_n)/R)^{1/2}} \underset{R\to\infty}{\overset{d}{\to}} N(0,1). \tag{2.11}$$

This holds for n finite and thus concerns the first moment of the small-n distribution of q_n. Writing $[\overrightarrow{\mathrm{Var}}_R(q_n)]^{1/2}$ as $\overrightarrow{\mathrm{SD}}_R(q_n)$ this implies that the interval

$$[\overrightarrow{E}_R(q_n) - 3\overrightarrow{\mathrm{SD}}_R(q_n)/\sqrt{R}, \overrightarrow{E}_R(q_n) + 3\overrightarrow{\mathrm{SD}}_R(q_n)/\sqrt{R}] \tag{2.12}$$

will contain the unknown $E(q_n)$ with a probability exceeding 99.5% for R sufficiently large.[1]

When the R elements of the series of $q_n^{(r)}$ have been collected in an EViews workfile then the standard descriptive statistics establish the MCS estimates $Mean = \overrightarrow{E}_R(q_n)$ and $Std.Dev. = \overrightarrow{\mathrm{SD}}_R(q_n)$, and with these the interval (2.12) for $E(q_n)$ is easily obtained. Because R can and should be chosen large, we can make $\overrightarrow{\mathrm{SD}}_R(q_n)/\sqrt{R}$ arbitrarily small, because $\mathrm{SD}(q_n)$ is of course invariant with respect to R. In that way a narrow confidence interval can be obtained, even when one chooses the confidence coefficient as large as at least 99.5%. Given that computing

[1] Calculation of the CDF of $N(0,1)$ at 3 by EViews gives @cnorm(3)=0.99865 and the 99.75% quantile of $N(0,1)$ is @qnorm(0.9975)=2.807.

time is almost free, the investigator can choose R such that there is no reason to stick to the habitual nominal significance level α of 5% and a corresponding confidence coefficient of 95%, when producing inference based on Monte Carlo estimates. One can afford a very small probability of type I errors and corresponding high coverage probability of confidence intervals, because by choosing R large, one can at the same time mitigate the width of confidence intervals. Therefore, unlike in applied econometrics, where sample sizes are mostly small and not under the control of the investigator, in the context of MCS we should usually employ a smaller significance level, for instance as small as about 0.5% by using 3 as a rule-of-thumb critical value for statistics that are asymptotically (large-R) standard normal.

Note that "the price" in terms of the required increase in the number of replications of augmenting the nominal confidence coefficient of an interval from 95% to 99.5%, and thus increasing the critical value from 1.96 to 2.81, while keeping the width of the confidence interval constant, is $(2.807/1.960)^2 = 2.05$. Thus roughly spoken, by doubling the number of replications we can either reduce the width of a confidence interval with given confidence coefficient to about $100/\sqrt{2}$ or 71% of its earlier length, or if it were a 95% confidence interval enhance its confidence coefficient to almost 99.5%.

How large R should be in general for the Normal approximation to be accurate for a sample average has been discussed in the preceding section. In (2.11) it is not just the behavior of the sample average that matters, but in fact that of its so-called Studentized transformation, in which in the denominator we have an estimated standard error and not the true standard deviation of the sample average. To examine the qualities of the normal approximation of a Studentized sample average we have adapted program msc15.prg as follows:

```
'mcs21.prg:  Studentized sample averages and the CLT
!R=1000
!sqrtR=@sqrt(!R)
workfile f:\MCS\mcs21.wf1 u 1 !R
!muv=1
!sigv=3
```

```
!a=5
!b=15
rndseed 9876543210
!metaR=10000
matrix (!metaR,5) simres
for !rep=1 to !metaR
   genr z=nrnd
   genr v=!muv + !sigv*z
   genr u=!a + (!b-!a)*rnd
   genr x=@rchisq(2)
   genr p=rnd>0.75
   genr w=p*(x+3) - (1-p)*(z^2+2)
   simres(!rep,1)=!sqrtR*@mean(z)/@stdev(z)
   simres(!rep,2)=!sqrtR*(@mean(v) - !muv)/@stdev(v)
   simres(!rep,3)=!sqrtR*(@mean(u) - (!a + !b)/2)/@stdev(u)
   simres(!rep,4)=!sqrtR*(@mean(x) - 2)/@stdev(x)
   simres(!rep,5)=!sqrtR*(@mean(w) +1)/@stdev(w)
next
simres.write f:\MCS\mcs21res.txt
workfile f:\MCS\mcs21res.wf1 u 1 !metaR
read f:\MCS\mcs21res.txt Studz Studv Studu Studx Studw
```

Note that the program actually performs a meta-simulation study. It examines the Studentized version of an estimator of an expectation $\overrightarrow{E}_R(q_n)$, where q_n is distributed as the random variables z, v, u, x, or w (hence n plays actually no role here). How close (2.11) gets for $R = 1{,}000$ replications to the standard normal distribution is investigated in a simulation with $metaR = 10{,}000$ replications. In these days of very fast computers it is seldom that a simulation study is based on less than 1,000 replications. In the earlier section we found out that for the nonnormal variables u, x, and w a sample size of 1000 seems reasonable to rely on the normal approximation for the sample average. Figure 2.1 presents similar results on Studentized sample averages.

Figure 2.1 shows that the normal approximation seems again quite reasonable, though not always perfect, for a sample size of 1,000. Note, however, that for the accuracy of the interval (2.12) it does not matter

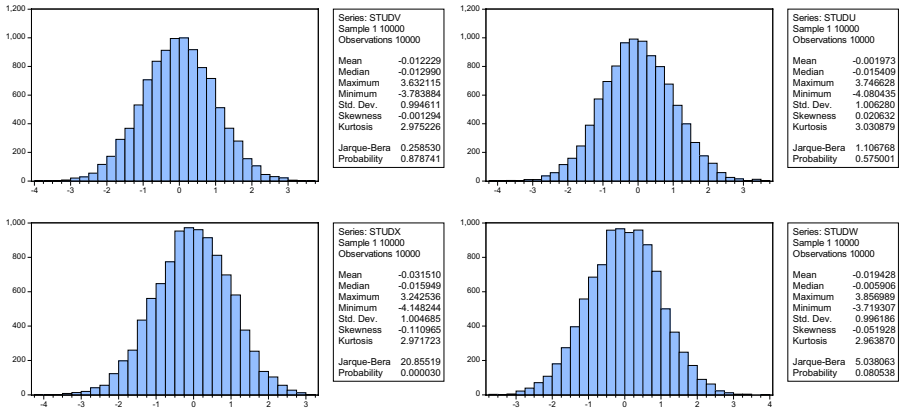

Fig. 2.1 Distribution of Studentized sample averages for $R = 1{,}000$ and $metaR = 10{,}000$.

whether the distribution conforms to the standard normal completely. The only thing that matters is whether the (two-sided) p-value at 3 is less than 0.5% and less than 5% at 2. From the histograms it is obvious that realizations in absolute value larger than 3 are quite rare. Estimating p-values from a MCS is one of the topics of the next section. Here we simply count how often the histograms in Figure 2.1 show outcomes that in absolute value exceed 3. This can be done in various ways in EViews. For instance (for the variable Studv), by typing in the EViews command window: series count=@cumsum(abs(Studu)>3). Next viewing series count shows in its last observation the required number. For Studv, Studu, Studx, and Studw these counts are: 20, 37, 28, and 19, respectively. Of course, these are random outcomes for which we will discuss their accuracy later, but they already indicate that when $R = 1{,}000$ a confidence interval for an expectation constructed with critical value 3 has an extremely high (most probably over 99.5%, but less than 100%) confidence coefficient. Thus, for random variables which do not deviate more severely from normal than u, x, and w do, taking sample averages at sample size $R = 1{,}000$ justifies reliance on the CLT's normality result. In what follows, however, we will demonstrate that for other reasons a Monte Carlo sample size of 1,000 may nevertheless be too small in relation to the (often implicit) claims made regarding Monte Carlo findings.

2.2 Analysis of Estimator Bias by Simulation

When q_n is actually an estimator $\hat{\beta}_j$ for some scalar parameter β_j from a vector β obtained from a sample of size n, then the above interval can be used to verify whether it could be the case (or not) that $\hat{\beta}_j$ is an unbiased estimator for β_j. Such inference can also be cast into the framework of a simple hypothesis test. The null-hypothesis $H_0 : E(\hat{\beta}_j) = \beta_j$ can be tested from the MCS sample by the MCS test statistic

$$\frac{\overrightarrow{E}_R(\hat{\beta}_j) - \beta_j}{\overrightarrow{\mathrm{SD}}_R(\hat{\beta}_j)/\sqrt{R}}. \tag{2.13}$$

Note the peculiar form of this null-hypothesis. We do not test $H_0 : \beta_j = c$ for some numerical value c by comparing $\hat{\beta}_j - c$ from a sample of size n of empirical data with $\widehat{SD}(\hat{\beta}_j) = [\widehat{\mathrm{Var}}(\hat{\beta}_j)]^{1/2}$, as we usually do in applied econometrics. But, exploiting that we know the numerical value that β_j has in the simulation, we test in a sample of Monte Carlo experiments of size R whether the first moment of $\hat{\beta}_j$, i.e., $E(\hat{\beta}_j)$ which is estimated by $\overrightarrow{E}_R(\hat{\beta}_j)$, equals this numerical value β_j. When the null-hypothesis $E(\hat{\beta}_j) = \beta_j$ is true and R large indeed, statistic (2.13) is approximately distributed as standard normal. Note that it is NOT distributed as Student with $n-k$ degrees of freedom, but approximately as Student with $R-1$ degrees of freedom (test for the mean in a sample of size R), which is practically equivalent to standard normal for R large.

As always, this test corresponds to checking whether or not β_j lies in an interval like (2.12). However, the interval provides much more information than the mere value of test statistic (2.13) does, because it communicates the test outcomes for the chosen significance level for any null hypothesis $H_0 : E(\hat{\beta}_j) = \beta_j^*$, where β_j^* is an arbitrary real value, and not just for $\beta_j^* = \beta_j$. At the chosen significance level the hypothesis $E(\hat{\beta}_j) = \beta_j^*$ is rejected for any β_j^* that is not in the interval, and otherwise that value is not rejected. Hence, the interval provides at the same time an indication of the magnitude of any bias, and therefore, as far as the MCS analysis of the expectation or bias of an econometric estimator is concerned, it seems preferable not to focus on the test statistic (2.13) and its significance or its p-value, but on constructing

a confidence interval with very high confidence coefficient, such as[2]

$$\Pr\{\overrightarrow{E}_R(\hat{\beta}_j) - 3[\overrightarrow{\mathrm{Var}}_R(\hat{\beta}_j)/R]^{1/2} \leq E(\hat{\beta}_j) \\ \leq \overrightarrow{E}_R(\hat{\beta}_j) + 3[\overrightarrow{\mathrm{Var}}_R(\hat{\beta}_j)/R]^{1/2}\} \overset{a}{\geq} 0.995$$

or, directly for the bias,

$$\Pr\{\overrightarrow{E}_R(\hat{\beta}_j) - \beta_j - 3\overrightarrow{SD}_R(\hat{\beta}_j)/\sqrt{R} \\ \leq E(\hat{\beta}_j) - \beta_j \leq \overrightarrow{E}_R(\hat{\beta}_j) - \beta_j + 3\overrightarrow{\mathrm{SD}}_R(\hat{\beta}_j)/\sqrt{R}\} \overset{a}{\geq} 0.995. \quad (2.14)$$

From the latter one can immediately assess (if zero is not in the interval) whether the unbiasedness hypothesis is rejected strongly (because the significance level is chosen very low) or not, and what (with high probability) the likely range is of the magnitude of the actual bias, if any.

Note that even when in an actual MCS study the focus is on some variant of the linear regression model, q_n does not necessarily represent an element of $\hat{\beta} = (X'X)^{-1}X'y$, the OLS estimator of the coefficient vector β in the model $y = X\beta + u$, where X is an $n \times k$ matrix. It could also pertain to the estimator $s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(n-k)$ of the disturbance variance σ_u^2, or be an element of the estimator of the variance matrix of the parameter estimates $\widehat{\mathrm{Var}}(\hat{\beta}) = s^2(X'X)^{-1}$. Then one might perhaps want to examine how much a particular element of $E[s^2(X'X)^{-1}]$ differs from the corresponding element of $n^{-1}\sigma_u^2 \mathrm{plim}_{n\to\infty}(X'X/n)^{-1}$.

Once again we want to underline, that MCS is not a method (like OLS or GMM) by which one can draw inferences on the actual values of the parameters of a model. In MCS a DGP is constructed and the parameter values are chosen by the investigator, and hence are completely known. From the simulation results inferences can be drawn on the qualities of particular statistical techniques (such as OLS), see the illustrations below. Such a quality might be unbiasedness. Note, though, that a MCS based test procedure on unbiasedness of an estimator can only corroborate or refute the hypothesis of unbiasedness

[2] Self-evidently, $\overset{a}{\geq}$ indicates here that the inequality is valid asymptotically for R large. Using not 3 but 2.81 as the critical value we could also have written $\overset{a}{=} 0.995$.

of q_n. It can of course never prove unbiasedness, because if the unbiasedness hypothesis is not rejected that does not imply that it is true, because we might commit a type II error. Also, it might indicate that the bias happens to be very small for the chosen parameter values, whereas it could be larger for different parameter values. Hence, if one cannot prove analytically that an estimator is unbiased, it doesn't seem to be of much practical use to test the unbiasedness hypothesis by a MCS study, except if one is eager to produce experimental evidence that the estimator is not unbiased. That type of inference and also an indication of the likely magnitude of the bias is directly available from an interval like (2.14), and therefore this is preferable to a single test outcome (2.13).

2.2.1 Choosing the MCS Sample Size

Let's have a further look into some of the issues relevant when choosing the value of R. Suppose that we wish to assess the bias of the estimator $q_n = \hat{\beta}_j$ of the coefficient β_j by simulation. The unbiased Monte Carlo estimate of this bias is $\overrightarrow{E}_R(\hat{\beta}_j) - \beta_j$. If we want to be pretty sure, say with a probability of at least about 99.5%, that the relative absolute error in this bias assessment, which we will address as the *precision* and denote by ϵ, will not exceed 100ϵ% of β_j, then this requires roughly, provided $\beta_j \neq 0$, that

$$\frac{3}{\sqrt{R}}[\mathrm{Var}(\hat{\beta}_j)]^{1/2} \leq \epsilon\beta_j \quad \text{or} \quad R \geq \frac{9}{\epsilon^2}\frac{\mathrm{Var}(\hat{\beta}_j)}{\beta_j^2}. \tag{2.15}$$

For $\hat{\beta}_j$ to be of some practical use for estimating β_j, it is generally required that $[\mathrm{Var}(\hat{\beta}_j)]^{1/2} \ll |\beta_j|$. So, for many situations it seems relatively safe to choose $R \approx 10/\epsilon^2$. So, when taking $R = 1{,}000$, we have $\epsilon \approx 0.1$ and the relative errors in the bias assessment will be smaller than 10% in general. For $R = 10{,}000$ they will be smaller than 3%, and smaller than 1% for $R = 100{,}000$. The actual accuracy will be much better if $[\mathrm{Var}(\hat{\beta}_j)]^{1/2} \ll |\beta_j|$ indeed. However, without information on this, for instance in the form of $\overrightarrow{\mathrm{Var}}_R(\hat{\beta}_j)/\beta_j^2$, no guarantees on the realized relative precision of $\overrightarrow{E}_R(\hat{\beta}_j) - \beta_j$ can be given.

Similarly if one wants to control the absolute error in the bias estimate, to be addressed as the *tolerance* τ. Then $3[\mathrm{Var}(\hat{\beta}_j)/R]^{1/2} \leq \tau$ leads roughly to $R \geq (10/\tau^2)\mathrm{Var}(\hat{\beta}_j)$. So, setting $\tau = 0.01$ requires $R \geq 1{,}000 \times \mathrm{Var}(\hat{\beta}_j)$. Hence, whether taking $R = 1{,}000$ or $R = 10{,}000$ is appropriate in a particular situation for achieving a certain required precision or tolerance does not just depend on the accuracy of the normal approximation to the distribution of a sample average, but also, and inevitably, on the actual magnitude of $\mathrm{Var}(\hat{\beta}_j)$.

Therefore, if one wants to be pretty sure beforehand about the realized value of the relative precision ϵ or the maximal absolute error τ of $\overrightarrow{E}_R(\hat{\beta}_j)$, one should first run a pilot Monte Carlo in which the major purpose is to assess $\overrightarrow{\mathrm{Var}}_R(\hat{\beta}_j)$, perhaps for all designs that will be examined (all relevant values of θ and n), and then choose a numerical upperbound for $\mathrm{Var}(\hat{\beta}_j)$, say $\overline{\mathrm{Var}}(\hat{\beta}_j)$. At the same time one can inspect the distribution of $\hat{\beta}_j$ to find out whether relying on the CLT sets further special requirements to the Monte Carlo sample size. If not, one could determine the required R exclusively by

$$R \geq \frac{9}{\epsilon^2}\frac{\overline{\mathrm{Var}}(\hat{\beta}_j)}{\beta_j^2} \quad \text{or} \quad R \geq \frac{9}{\tau^2}\overline{\mathrm{Var}}(\hat{\beta}_j). \tag{2.16}$$

In practice, primary interest usually does not focus just on the moment $E(\hat{\beta}_j)$ or the bias, but also on $\mathrm{Var}(\hat{\beta}_j)$ for its own sake; not just as a determining factor of the precision or tolerance of $\overrightarrow{E}_R(\hat{\beta}_j)$. The qualities of different possibly biased estimators are usually compared on the basis of their MSE (mean squared error). This depends on both $E(\hat{\beta}_j)$ and $\mathrm{Var}(\hat{\beta}_j)$ and therefore it is not just the precision of $\overrightarrow{E}_R(\hat{\beta}_j)$, which depends on the value of $\mathrm{Var}(\hat{\beta}_j)$, that we should bother about, but also the precision of $\overrightarrow{\mathrm{Var}}_R(\hat{\beta}_j)$ as such, which we will address after a first simple illustration of MCS.

2.2.2 MCS of a Standard Normal Linear Regression Model

The theory on assessing an expectation as given above will first be applied to a model where simulation is actually not required, because all properties of OLS can be obtained here from simple exact finite sample analytic derivations. However, the advantage is that this example allows to illustrate how close (or not close at all) one can get to the truth by

MCS approximations. The DGP has particular numerical values for the parameters and the single regressor variable of the simple classic normal linear regression model

$$y_i = \beta_1 + \beta_2 x_i + u_i, \quad u_i \sim NIID(0, \sigma^2), \quad i = 1, \ldots, n. \tag{2.17}$$

We use the program:

```
'mcs22.prg: MCS of a standard normal linear regression
!n=60                              'sample size of econometric model
workfile f:\MCS\mcs21.wf1 u 1 !n
genr i=@trend(0)                   'variable i has values 1,...,n
genr x=i-20*@floor((i-1)/20) 'variable x has values 1,...,20,1,...,20,1,...,20
!beta1=5                           'true value of intercept
!beta2=0.8                         'true value of slope coefficient
!sigma=1.5                         'true value of disturbance stand. deviation
genr Xbeta=!beta1+!beta2*x         'deterministic component of dep. variable
rndseed 9876543210
!R=10000                           'sample size of MCS
matrix (!R,4) simres               'declaration of a !Rx4 matrix simres
for !rep=1 to !R                   'loop of the replications
   genr usim=!sigma*nrnd           'generation vector of IID N[0,!sigma^2]
   genr ysim=Xbeta + usim          'generation of dep. var. N(Xbeta,!sigma^2*I)
   equation eq1.ls ysim=c(1)+c(2)*x   'OLS of simple regression model
   simres(!rep,1)=eq1.@coefs(1)    'all c(1) in 1st column simres
   simres(!rep,2)=eq1.@coefs(2)    'all c(2) in 2nd column simres
   simres(!rep,3)=eq1.@se          'OLS estimate of sigma in 3rd
   simres(!rep,4)=eq1.@stderrs(2)  'StdErr(b2) collected in 4th column
next                               'now all !R replications completed
simres.write f:\MCS\mcs22sim.txt   'simres saved as tex file
workfile f:\MCS\mcs22sim.wf1 u 1 !R  'workfile size !R
read f:\MCS\mcs22sim.txt b1 b2 s b2se  'simres written on 2nd wf
                                          'giving names: b1 b2 s en b2se
genr s2=s*s                        's2 contains estimates of !sigma^2
genr b2var=b2se^2                  'b2var contains estimates of Var(b2)
```

After running this program workfile mcs22.wf1 contains in variable usim the R^{th} replication of the $n \times 1$ vector of disturbances and in ysim the R^{th} replication of the vector of dependent variables. In eq1 the results of the R^{th} application of OLS (on a sample of $n = 60$ observations) are still available (all earlier replications have been overwritten). These OLS results are:

```
Dependent Variable: YSIM
Method: Least Squares
Sample: 1 60
```

```
Included observations: 60
YSIM=C(1)+ C(2)*X

Variable  Coefficient  Std. Error  t-Statistic  Prob.
C(1)       5.594788     0.349917    15.98889   0.0000
C(2)       0.764225     0.029211    26.16268   0.0000

R-squared           0.921884  Mean dependent var     13.61915
Adjusted R-squared  0.920537  S.D. dependent var     4.628378
S.E. of regression  1.304700  Akaike info criterion  3.402589
Sum squared resid   98.73005  Schwarz criterion      3.472400
```

Note that the results in eq1 are according to expectation for an adequately specified regression model with true intercept 5 and true slope 0.8, because the OLS estimates do not deviate significantly from these values. The S.E. of regression (s) seems pretty large, given the value of σ, which is 1.5. Some graphical results can be found in Figure 2.2. It shows that the 10,000th vector of disturbances happen to contain some

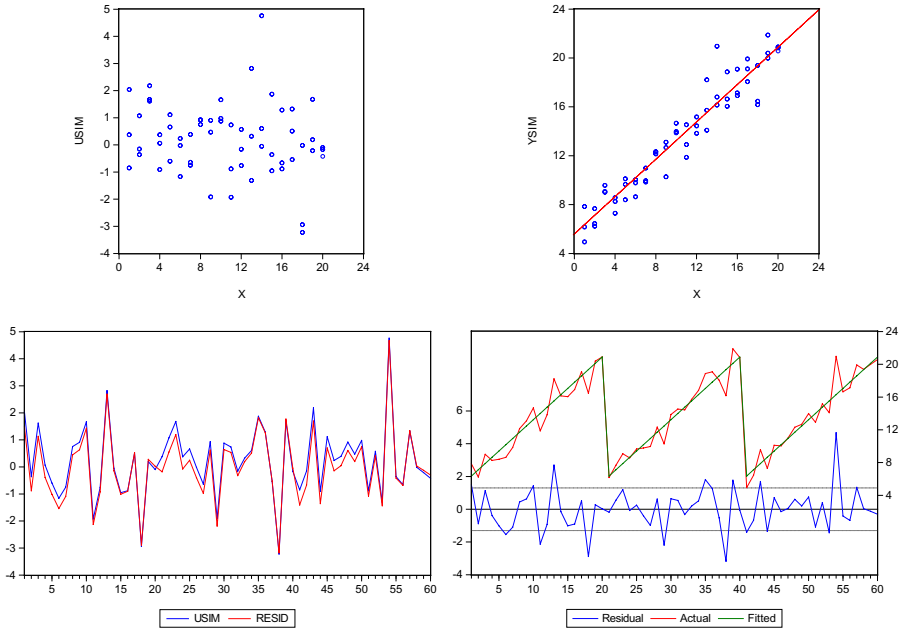

Fig. 2.2 Results of the 10,000th replication of OLS estimation in mcs22.prg.

quite large realizations (explaining the large value of s). However, to depict the general OLS properties, we should not just look at one arbitrary drawing of the $n \times 1$ random vector of disturbances u, giving rise to one arbitrary realization of the dependent variable $y = X\beta + u$, from which one realization has been obtained for the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$, the residual vector $\hat{u} = y - X\hat{\beta}$, the statistic $s^2 = \hat{u}'\hat{u}/(n-k)$ with $n - k = 58$ and the matrix $s^2(X'X)^{-1}$ which is used for estimating $\mathrm{Var}(\hat{\beta})$.

By examining not just 1 but all 10,000 simulated realizations of $\hat{\beta}$, s^2, and $s^2(X'X)^{-1}$ we can establish their major statistical properties. Workfile mcs22sim.wf1 contains all the stored MCS results. Figure 2.3 presents the histograms of the various simulated statistics. From the histogram of B1, which collects all realizations of the OLS estimator of the intercept $\hat{\beta}_1$, we find[3] $\overrightarrow{E}(\hat{\beta}_1) = 5.004656$

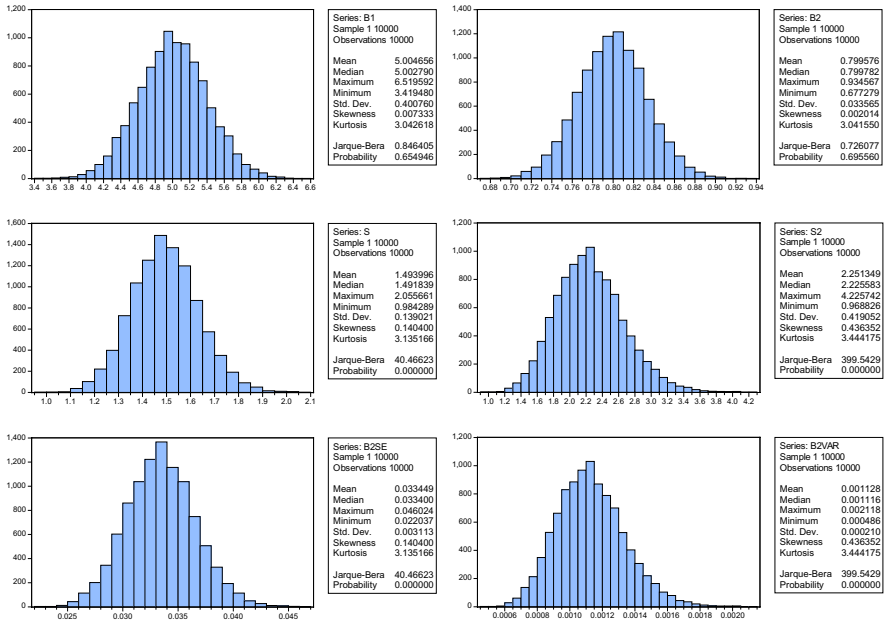

Fig. 2.3 MSC results of program mcs22.prg.

[3] Like we do not specify n in $\hat{\beta}_1$, whereas we did in q_n, we will no longer now specify R all the time in Monte Carlo estimates, because it is self-evident that all MCS estimators $\overrightarrow{E}$ and $\overrightarrow{\mathrm{Var}}$ are obtained from a sample of R drawings.

and $\overrightarrow{\mathrm{Var}}(\overrightarrow{E}(\hat{\beta}_1)) = 0.404472^2/10{,}000$ or $\overrightarrow{SD}(\overrightarrow{E}(\hat{\beta}_1)) = 0.00404$. So, we immediately see that, because $0.004656/0.00404 \ll 3$, we do not have to reject the hypothesis $E(\hat{\beta}_1) = 5$, which does not surprise, because we know that OLS is unbiased in the classic (which means that all regressors are exogenous) linear regression model. That for variable B1 Skewness is close to zero and Kurtosis close to 3, so that the Jarque–Bera statistic does not have a small p-value, is plausible too, because we know that the OLS estimator $\hat{\beta}$ follows a multivariate normal distribution in the classic linear regression model when the disturbances are normal. For the very same reasons the results in the histogram of B2 are plausible too.

We know that s^2 is unbiased for σ^2. We find for the MCS estimate of its bias $\overrightarrow{E}(s^2) - \sigma^2 = 0.001349$. Whether or not this is small cannot be judged on the basis of this actual value, but only by comparing it with the estimate of its standard error. For this we find $\overrightarrow{\mathrm{SD}}[\overrightarrow{E}(s^2)] = 0.0042$, giving a test statistic for testing the hypothesis of unbiasedness of $0.001349/0.0042$ and again this does not lead to rejection, although this could have happened; after all, MCS results are not exact but random, and even at sample size 10,000 the probability to commit a type I error follows the significance level used.

Note that in the normal classic model we have $\hat{u}'\hat{u}/\sigma^2 \sim \chi^2_{n-k}$. Hence, $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$ and that distribution, and hence s^2 itself, has Skewness $2\sqrt{2/(n-k)}$ and Kurtosis $3 + 12/(n-k)$, which for $n - k = 58$ amount to 0.371 and 3.207, respectively. These values do not seem to match very well with the MCS estimates found for S2, although they do reject normality. Below it will become clear that a MCS sample size of 10,000 may usually be sufficient to accurately estimate first moments, but will often be much too small for the accurate assessment of higher-order moments.

Tease your brains by interpreting the histogram of B2SE and comparing its mean with the Std. Dev. mentioned next to the histogram of B2. OLS estimator $\hat{\beta}_2$ is random (due to the randomness of vector u). Its true variance (standard deviation) can be obtained from $\sigma^2(X'X)^{-1}$. It can be calculated (see Exercise 6) to be 0.0011278 with square-root 0.03358. A MCS assessment of this standard deviation is given next to the histogram of B2 and amounts to 0.033565, yielding

$\overrightarrow{\text{Var}}(\hat{\beta}_2) = 0.033565^2 = 0.0011266$ for the (unbiased) MCS estimate of the appropriate element of $\sigma^2(X'X)^{-1}$. A practitioner does not know σ^2 but obtains from just one sample of size n an estimate s^2 and uses $\widehat{\text{Var}}(\hat{\beta}) = s^2(X'X)^{-1}$. Histogram B2VAR shows how much $\widehat{\text{Var}}(\hat{\beta}_2)$ can vary from sample to sample. From this histogram we can also obtain a confidence interval for $E[\widehat{\text{Var}}(\hat{\beta}_2)] = \text{Var}(\hat{\beta}_2)$ with very high confidence coefficient, namely

$$0.001128 \pm 3 * 0.00000210 \Rightarrow [0.001122\ ,\ 0.001134], \tag{2.18}$$

which (as expected) contains the true value 0.0011278 indeed. However, the fact that this interval also contains the unbiased MCS estimate $\overrightarrow{\text{Var}}(\hat{\beta}_2) = 0.0011266$, which we obtained from the histogram of B2, is not self-evident. The explanation for that is that this is an estimator, thus random, and hence the magnitude of its standard error, i.e., $\overrightarrow{\text{SD}}[\overrightarrow{\text{Var}}(\hat{\beta}_2)]$, should explain by how much it may differ from $\text{Var}(\hat{\beta}_2) = 0.0011278$, and not $\overrightarrow{\text{SD}}[\widehat{\text{Var}}(\hat{\beta}_2)]$ which determines the width of interval (2.18). The assessment of notions like $\overrightarrow{\text{SD}}[\overrightarrow{Var}(\hat{\beta}_2)]$ is discussed in the next section.

Here, where we do know the true value of $E(\hat{\beta}_2)$, it is easy to check that the true MCS error made in assessing the bias, which is $0.799576 - 0.8 = -0.000424$, has an absolute value which is a lot smaller than τ, which is here $3 \times \text{SD}(\hat{\beta}_2)/\sqrt{R} \approx 0.001$. For a different value of the seed we might have found a much less accurate bias assessment, but it would be highly improbable to find an error that exceeds τ.

The above illustrates that, even when we run a reasonably high number of replications, the inherent randomness of MCS results does never allow to arrive at positive inferences such as unbiasedness. We will see in what follows that MCS is more suited to signal finite sample problems such as bias or nonnormality of estimators and size distortions of test procedures, when present.

2.3 Assessing the (R)MSE of an Estimator

When analyzing by MCS the properties of an estimator, say $\hat{\beta}_j$ with true value β_j, we are usually not only interested in its expectation $E(\hat{\beta}_j)$ and first centered moment or mean error, the bias $E(\hat{\beta}_j - \beta_j)$, but also

in its second centered moment, its variance $\mathrm{Var}(\hat{\beta}_j) = E[\hat{\beta}_j - E(\hat{\beta}_j)]^2$, assuming that these exist. Not only is the latter an indicator of the magnitude of the estimation errors when the estimator is unbiased. But, as we learned above, $\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)$ also provides the key to analyze the accuracy of $\overrightarrow{E}(\hat{\beta}_j)$ and of $\overrightarrow{E}(\hat{\beta}_j - \beta_j)$.

When the estimator is not unbiased an indicator of the magnitude of its estimation errors is provided by its MSE (mean squared error), or its square root the RMSE, where $\mathrm{MSE}(\hat{\beta}_j) \equiv E(\hat{\beta}_j - \beta_j)^2$. For an unbiased estimator the MSE specializes to the variance and the RMSE to the SD. Natural MCS estimators are

$$\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j) \equiv \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_j^{(r)} - \beta_j)^2 \quad \text{and} \quad \overrightarrow{\mathrm{RMSE}}(\hat{\beta}_j) \equiv [\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j)]^{1/2}, \tag{2.19}$$

where the first is unbiased. Because

$$\begin{aligned}
\overrightarrow{\mathrm{Var}}(\hat{\beta}_j) + [\overrightarrow{E}(\hat{\beta}_j) - \beta_j]^2 &= \frac{1}{R-1}\sum_{r=1}^{R}[\hat{\beta}_j^{(r)} - \overrightarrow{E}(\hat{\beta}_j)]^2 + [\overrightarrow{E}(\hat{\beta}_j) - \beta_j]^2 \\
&\underset{R\to\infty}{\overset{a}{=}} \frac{1}{R}\sum_{r=1}^{R}[\hat{\beta}_j^{(r)} - \overrightarrow{E}(\hat{\beta}_j)]^2 + [\overrightarrow{E}(\hat{\beta}_j) - \beta_j]^2 \\
&= \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_j^{(r)})^2 - 2\beta_j\overrightarrow{E}(\hat{\beta}_j) + \beta_j^2 = \overrightarrow{\mathrm{MSE}}(\hat{\beta}_j),
\end{aligned}$$

$\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j)$ can either be assessed directly or, when R is large, from the simulation results on $\overrightarrow{E}(\hat{\beta}_j)$ and $\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)$.

Especially when alternative not necessarily unbiased estimation techniques for a parameter β_j are being compared by simulation, say $\hat{\beta}_j$ and $\hat{\beta}_j^*$, one often ranks them according to the outcomes of the MCS estimates $\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j)$ and $\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j^*)$. In doing so, there are problems regarding the randomness of these MCS estimates. Here we pay attention to MCS estimation errors in separate (R)MSE assessments. Later (in Section 5.4) we will address the possible in(ter)dependence of $\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j)$ and $\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j^*)$ and also of $\overrightarrow{E}(\hat{\beta}_j)$ and $\overrightarrow{E}(\hat{\beta}_j^*)$.

In most simulation studies very little attention is being paid to the possible magnitude of the estimation errors (τ and ϵ) in $\overrightarrow{E}(\hat{\beta}_j)$,

and so it should not surprise that those in $\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)$ and $\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j)$ are hardly ever discussed at all. By generalizing in the next subsection the approach regarding the accuracy of the MCS estimate of the first moment, we will find that a relative precision of second moment estimates similar to that of the first moment may require substantially larger values of R.

2.3.1 Accuracy of MCS Estimates of Variance and Standard Deviation

For the IID sample $q_n^{(1)}, \ldots, q_n^{(R)}$ we assume that the uncentered moments $\mu_l \equiv E(q_n^{(r)})^l$ exist for positive integer l up to a certain appropriate value. We also define the centered moments $\mu_l^c \equiv E(q_n^{(r)} - \mu_1)^l$. Hence, μ_1^c is the bias and μ_2^c is the variance. Moreover, $\lambda \equiv \mu_3^c/(\mu_2^c)^{3/2}$ is the skewness and $\kappa \equiv \mu_4^c/(\mu_2^c)^2$ is the kurtosis of q_n. Obvious MCS estimators are

$$\overrightarrow{\mu}_l \equiv \frac{1}{R}\sum_{r=1}^{R}(q_n^{(r)})^l \quad \text{and} \quad \overrightarrow{\mu}_l^c \equiv \frac{1}{R-1}\sum_{r=1}^{R}(q_n^{(r)} - \overrightarrow{\mu}_1)^l. \tag{2.20}$$

According to Cramér (1946, p. 365), we have

$$\sqrt{R}(\overrightarrow{\mu}_2^c - \mu_2^c) \underset{R\to\infty}{\overset{d}{\to}} N(0, \mu_4^c - (\mu_2^c)^2), \tag{2.21}$$

where we can rewrite $\mu_4^c - (\mu_2^c)^2 = (\kappa - 1)(\mu_2^c)^2$. This suggests to estimate the variance of the MCS estimate $\overrightarrow{\mu}_2^c$ of the variance μ_2^c by

$$\overrightarrow{\mathrm{Var}}(\overrightarrow{\mu}_2^c) = \frac{(\overrightarrow{\kappa} - 1)(\overrightarrow{\mu}_2^c)^2}{R}. \tag{2.22}$$

By the delta-method (see Appendix B) one easily obtains from this for the MCS estimate of a standard deviation its estimated variance

$$\overrightarrow{\mathrm{Var}}((\overrightarrow{\mu}_2^c)^{1/2}) = \frac{(\overrightarrow{\kappa} - 1)}{4}\frac{\overrightarrow{\mu}_2^c}{R}. \tag{2.23}$$

From these it follows that, for R large and a confidence coefficient exceeding 99.5%, a confidence interval for the variance μ_2^c around $\overrightarrow{\mu}_2^c$ has half-width $3\overrightarrow{\mu}_2^c[(\overrightarrow{\kappa} - 1)/R]^{1/2}$, and similarly the interval for $(\mu_2^c)^{1/2}$ has half-width of $1.5[(\overrightarrow{\kappa} - 1)\overrightarrow{\mu}_2^c/R]^{1/2}$ around $(\overrightarrow{\mu}_2^c)^{1/2}$. Thus,

the relative error ϵ in a MCS variance estimate will usually not exceed $3[(\kappa - 1)/R]^{1/2}$. So, if the kurtosis of q_n is not too far from normal, say it does not exceed 5, then ϵ does not exceed $6R^{-1/2}$, implying that variance estimates with relative errors below 1% may require $R \geq 360{,}000$, whereas when $R = 1{,}000$ the relative error is with 99.5% confidence not exceeding 20%. For an estimated standard error the ϵ will as a rule not exceed $1.5[(\kappa - 1)/R]^{1/2}$ so that (for similar upperbound regarding κ) errors below 1% require $R \geq 180{,}000$, whereas for $R = 1{,}000$ the relative errors might be as big as 9.5%.

Thus, we find that often a much larger number of replications will be required for precise MCS estimation of a variance and its square root than for equally accurate estimation of an expectation. It is interesting to examine the implications of the above for some of the variance estimates obtained earlier. We focus on the results presented in Figure 1.5 and some of its consequences, all collected in Table 2.1.

For the four distinct variables of Figure 1.5, Table 2.1 presents in the rows $\overrightarrow{\mathrm{SD}}(\cdot)$ and $\overrightarrow{\kappa}(\cdot)$ the figures on the MCS estimates of their standard deviation and their kurtosis. The rows $\mathrm{Var}(\cdot)$ and $\mathrm{SD}(\cdot)$ contain the true values of the variance and the standard deviation of the four variables. The row $\overrightarrow{\mathrm{Var}}(\cdot)$ is simply the square of $\overrightarrow{\mathrm{SD}}(\cdot)$. The difference between $\overrightarrow{\mathrm{Var}}(\cdot)$ and $\mathrm{Var}(\cdot)$ is the MCS estimation error in the variance, which has expectation zero. The actual estimation errors in these 10,000 replications do have a standard deviation which is estimated from the MCS according to (2.22) by

$$\overrightarrow{\mathrm{SD}}[\overrightarrow{\mathrm{Var}}(\cdot)] = \overrightarrow{\mathrm{Var}}(\cdot)[(\overrightarrow{\kappa}(\cdot) - 1)/R]^{1/2}. \tag{2.24}$$

Table 2.1. MCS estimates of the SD and Var of IID variables and of their SDs.

	Variable			
	v	u	x	w
$\overrightarrow{\mathrm{SD}}(\cdot)$	3.0234	2.8915	1.9752	3.7991
$\mathrm{SD}(\cdot)$	3.0000	2.8868	2.0000	3.8079
$\overrightarrow{\mathrm{Var}}(\cdot)$	9.1409	8.3608	3.9014	14.4332
$\mathrm{Var}(\cdot)$	9.0000	8.3333	4.0000	14.5000
$\overrightarrow{\kappa}(\cdot)$	2.9670	1.7921	8.2000	3.7737
$\overrightarrow{\mathrm{SD}}[\overrightarrow{\mathrm{Var}}(\cdot)]$	0.1282	0.0744	0.1047	0.2404
$\overrightarrow{\mathrm{SD}}[\overrightarrow{\mathrm{SD}}(\cdot)]$	0.0212	0.0129	0.0265	0.0316

Note that the actual MCS estimation errors in the variance have an absolute value which is always smaller than twice the MCS estimate of their standard deviation, which is plausible. The difference between $\overrightarrow{\mathrm{SD}}(\cdot)$ and $\mathrm{SD}(\cdot)$ has an expectation converging toward zero for R large. The absolute value of this difference is also of a plausible magnitude when compared with the MCS estimate of its standard deviation, which is obtained according to (2.23) by

$$\overrightarrow{\mathrm{SD}}[\overrightarrow{\mathrm{SD}}(\cdot)] = \frac{1}{2}\overrightarrow{\mathrm{SD}}(\cdot)[(\overrightarrow{\kappa}(\cdot) - 1)/R]^{1/2}. \tag{2.25}$$

2.3.2 Accuracy of MCS Estimates of a (R)MSE

For estimators that are not unbiased we are interested in assessing their (R)MSE. Denoting the true scalar value which we try to estimate by q_n as ϑ, then the MSE can be represented as $\mu_2^{\vartheta} \equiv E(q_n - \vartheta)^2$ and its unbiased MCS estimator is

$$\overrightarrow{\mu}{}_2^{\vartheta} \equiv \frac{1}{R}\sum_{r=1}^{R}(q_n^{(r)} - \vartheta)^2. \tag{2.26}$$

Evaluating the variance of the squared error (VSE) of q_n gives

$$\begin{aligned}
\mathrm{VSE}(q_n) &\equiv \mathrm{Var}[(q_n - \vartheta)^2] = E(q_n - \vartheta)^4 - [E(q_n - \vartheta)^2]^2 \\
&= E[(q_n - \mu_1) + (\mu_1 - \vartheta)]^4 - \{E[(q_n - \mu_1) + (\mu_1 - \vartheta)]^2\}^2 \\
&= (\kappa - 1)(\mu_2^c)^2 + 4\lambda(\mu_1 - \vartheta)(\mu_2^c)^{3/2} + 4(\mu_1 - \vartheta)^2\mu_2^c,
\end{aligned} \tag{2.27}$$

where we used $\mu_1^c = 0$. Next, a standard CLT easily yields

$$\sqrt{R}(\overrightarrow{\mu}{}_2^{\vartheta} - \mu_2^{\vartheta}) \underset{R\to\infty}{\overset{d}{\to}} N(0, \mathrm{VSE}(q_n)). \tag{2.28}$$

From this we have

$$\begin{aligned}
\overrightarrow{\mathrm{SD}}(\overrightarrow{\mu}{}_2^{\vartheta}) &= R^{-1/2}[\overrightarrow{\mathrm{VSE}}(q_n)]^{1/2} \\
&= R^{-1/2}\{(\overrightarrow{\kappa} - 1)(\overrightarrow{\mu}{}_2^c)^2 + 4\overrightarrow{\lambda}(\overrightarrow{\mu}{}_1 - \vartheta)(\overrightarrow{\mu}{}_2^c)^{3/2} \\
&\quad + 4(\overrightarrow{\mu}{}_1 - \vartheta)^2\overrightarrow{\mu}{}_2^c\}^{1/2}
\end{aligned} \tag{2.29}$$

and three times this value gives the half width of the confidence interval for $\mu_2^{\vartheta} = MSE(q_n)$ around $\overrightarrow{\mu}{}_2^{\vartheta}$. By the delta-method one easily obtains

for the MCS estimate of a RMSE, denoted as $(\overrightarrow{\mu}_2^{\vartheta})^{1/2}$, that its standard deviation can be estimated by

$$\overrightarrow{\mathrm{SD}}[(\overrightarrow{\mu}_2^{\vartheta})^{1/2}] = \frac{1}{2}\left\{\frac{(\overrightarrow{\kappa}-1)(\overrightarrow{\mu}_2^{c})^2 + 4\overrightarrow{\lambda}(\overrightarrow{\mu}_1-\vartheta)(\overrightarrow{\mu}_2^{c})^{3/2} + 4(\overrightarrow{\mu}_1-\vartheta)^2\overrightarrow{\mu}_2^{c}}{R\overrightarrow{\mu}_2^{\vartheta}}\right\}^{1/2}. \tag{2.30}$$

The last two results provide the keys to express, or to control by adapting R, the accuracy of (R)MSE estimates, as is summarized in the next subsection.

2.4 Summary on Accuracy of MCS Moment Estimates

When in doubt whether in a particular situation a chosen R might lead to unacceptably high MCS estimation errors in the moment estimates $\overrightarrow{E}(\hat{\beta}_j)$, $\overrightarrow{\mathrm{SD}}(\hat{\beta}_j)$, and/or $\overrightarrow{\mathrm{RMSE}}(\hat{\beta}_j)$, or the squares of the latter two, an easy and pragmatic approach would be the following. One could run the program a few times for this number of replications using different seed values, and then examine by how much the MCS estimates vary. If the variation is acceptably small then this is reassuring evidence that the chosen value of R is sufficiently large. If it is found, however, that the variability of the outcomes is too large, and if one wants to decrease it by a factor f then one should increase the number of replications by a factor f^2.

Here we summarize a more sophisticated and guaranteed approach regarding how to choose R when one wants to achieve a certain tolerance (τ) or precision (ϵ) with very high probability. Its formulas also provide the keys to the disclosure of the actual precision of reported simulation results on moment estimates of parameter estimates, provided that next to β_j and R they contain the MCS estimates $\overrightarrow{E}(\hat{\beta}_j)$ and/or $\overrightarrow{E}(\hat{\beta}_j-\beta_j)$, and also $\overrightarrow{\mathrm{Var}}(\hat{\beta}_j) = [\overrightarrow{SD}(\hat{\beta}_j)]^2$ and/or $\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j) = [\overrightarrow{\mathrm{RMSE}}(\hat{\beta}_j)]^2$, and ideally as well the skewness $\overrightarrow{\lambda}(\hat{\beta}_j)$ and the kurtosis $\overrightarrow{\kappa}(\hat{\beta}_j)$. If the latter two are not available one should choose relatively safe upperbound values for them, say 2 for the skewness if $\hat{\beta}_j$ has a positive bias and -2 otherwise, and perhaps 5 for the kurtosis, like suggested before.

Assuming that the distribution of $\hat{\beta}_j$ and the value of R are such that the critical value 3 guarantees a very high confidence coefficient for the intervals on $E(\hat{\beta}_j)$, $\mathrm{SD}(\hat{\beta}_j)$, $\mathrm{Var}(\hat{\beta}_j)$, $\mathrm{RMSE}(\hat{\beta}_j)$ and $\mathrm{MSE}(\hat{\beta}_j)$, very safe upperbounds to these errors can be estimated from the MCS (and therefore they carry the MCS arrow) using the formulas given below. They follow directly from the above derivations (but ϑ is now replaced by β_j).

With respect to $\overrightarrow{E}(\hat{\beta}_j)$ or $\overrightarrow{E}(\hat{\beta}_j - \beta_j)$:

$$\overrightarrow{\tau} = \frac{3}{\sqrt{R}}\overrightarrow{\mathrm{SD}}(\hat{\beta}_j) = \frac{3}{\sqrt{R}}\{[\overrightarrow{\mathrm{RMSE}}(\hat{\beta}_j)]^2 - [\overrightarrow{E}(\hat{\beta}_j - \beta_j)]^2\}^{1/2},$$
$$\overrightarrow{\epsilon} = \overrightarrow{\tau}/\beta_j \quad \text{if } \beta_j \neq 0. \tag{2.31}$$

With respect to $\overrightarrow{\mathrm{SD}}(\hat{\beta}_j)$:

$$\overrightarrow{\epsilon} = 1.5[(\overrightarrow{\kappa}(\hat{\beta}_j) - 1)/R]^{1/2},\ \overrightarrow{\tau} = \overrightarrow{\epsilon} \times \overrightarrow{SD}(\hat{\beta}_j). \tag{2.32}$$

With respect to $\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)$:

$$\overrightarrow{\epsilon} = 3[(\overrightarrow{\kappa}(\hat{\beta}_j) - 1)/R]^{1/2},\ \overrightarrow{\tau} = \overrightarrow{\epsilon} \times \overrightarrow{\mathrm{Var}}(\hat{\beta}_j). \tag{2.33}$$

With respect to $\overrightarrow{\mathrm{RMSE}}(\hat{\beta}_j)$:

$$\overrightarrow{\tau} = \frac{1.5/\sqrt{R}}{\overrightarrow{\mathrm{RMSE}}(\hat{\beta}_j)}\{[\overrightarrow{\kappa}(\hat{\beta}_j) - 1][\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)]^2 + 4\overrightarrow{\lambda}(\hat{\beta}_j)[\overrightarrow{E}(\hat{\beta}_j - \beta_j)]$$
$$\times[\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)]^{3/2} + 4[\overrightarrow{E}(\hat{\beta}_j) - \beta_j)]^2\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)\}^{1/2}, \tag{2.34}$$
$$\overrightarrow{\epsilon} = \overrightarrow{\tau}/\overrightarrow{\mathrm{RMSE}}(\hat{\beta}_j).$$

With respect to $\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j)$:

$$\overrightarrow{\tau} = \frac{3}{\sqrt{R}}\{[\overrightarrow{\kappa}(\hat{\beta}_j) - 1][\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)]^2 + 4\overrightarrow{\lambda}(\hat{\beta}_j)[\overrightarrow{E}(\hat{\beta}_j - \beta_j)][\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)]^{3/2}$$
$$+4[\overrightarrow{E}(\hat{\beta}_j) - \beta_j)]^2\overrightarrow{\mathrm{Var}}(\hat{\beta}_j)\}^{1/2}, \tag{2.35}$$
$$\overrightarrow{\epsilon} = \overrightarrow{\tau}/\overrightarrow{\mathrm{MSE}}(\hat{\beta}_j).$$

Note that when $\overrightarrow{E}(\hat{\beta}_j - \beta_j) = 0$ the (R)MSE results simplify to the SD/VAR results, which is logical.

Replacing in the above formulas $\overrightarrow{\tau}$ by τ and $\overrightarrow{\epsilon}$ by ϵ and reformulating them such that they are explicit in R indicates how R should be

chosen to guarantee particular values of τ and ϵ regarding either the first moment, the (square root of the) second centered moment or the (R)MSE.

2.5 Moments of OLS in Stable AR(1) Models

To illustrate the above we now consider a situation where no simple exact analytic solution exists regarding the bias of the OLS estimator, namely in dynamic regression models. We focus on the AR(1) model with intercept, where

$$y_t = \beta_1 + \beta_2 y_{t-1} + u_t. \tag{2.36}$$

Of course, a full description of the DGP requires the actual numerical values for the parameters, a specification of y_0 and also of the distribution of u. Initially we shall look at the very specific case where $\beta_1 = 0$, $\beta_2 = 0.5$, $y_0 = 0$, and $u_t \sim NIID(0,1)$, see program mcs23.prg.

```
'mcs23.prg: MCS of an AR(1) model
!R=100               'number of Monte Carlo replications
!n=25                'sample size of the regressions
workfile f:\MCS\mcs23.wf1 u 0 !n
!beta1=0
!beta2=0.5
rndseed 9876543210
smpl 0 0
genr y=0
      'genr y=!beta1/(1-!beta2)
      'genr y=!beta1/(1-!beta2) + nrnd/@sqrt(1-!beta2^2)
smpl 1 !n
matrix (!R,2) simres
for !rep=1 to !R
  genr u = nrnd
  genr y = !beta1 + !beta2*y(-1) + u
  equation eq1.ls y c y(-1)
  simres(!rep,1)=eq1.@coefs(1)
  simres(!rep,2)=eq1.@coefs(2)
next
simres.write f:\MCS\mcs23sim.txt
workfile f:\MCS\mcs23sim.wf1 u 1 !R
read f:\MCS\mcs23sim.txt b1 b2  'names simulated var's: b1, b2
```

Note that both $!n$ and $!R$ are pretty small, which may give rise to serious finite sample problems of the (asymptotically valid) econometric techniques, and as well a poor registration of these due to lack of precision of the MCS study. However, we will compare these with results for larger values of n and R to illustrate their different roles and effects.

Although we fix the initial value $y_0 = 0$, it will be easy to change the program later such that either $y_0 = \beta_1/(1-\beta_2)$ or $y_0 \sim N(\beta_1/(1-\beta_2), 1-\beta_2^2)$, simply by moving the comment symbol $'$. After the $R \times 2$ matrix with the name simres (again) is defined, we go through a loop in which the R simulation replications are performed. In these, for $rep = 1, \ldots, R$, each time a fresh (independent) series of n *NIID*(0,1) disturbances is generated in u, and for $t = 1, \ldots, n$ the observations $y_t = \beta_1 + \beta_2 y_{t-1} + u_t$ are obtained recursively. Then, in eq1 the regression results of LS y c y(−1) are stored, and in particular we put $simres(rep, 1) = \hat{\beta}_1$ and $simres(rep, 2) = \hat{\beta}_2$. The final three lines of the program produce another EViews workfile, namely mcs23sim.wf1, which contains 2 variables with names b1 and b2. These variables establish the Monte Carlo sample consisting of variables of R (hence NOT n) observations, namely the R random and IID drawings from the distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. Note that all the R coefficient estimates are each obtained from regressions with sample size n.

We run the program and obtain the following two histograms for the variables b1 and b2 of workfile mcs23sim.wf1, respectively.

Note that from $R = 100$ we do not get a clear picture on the population distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$ yet in Figure 2.4. For both no significant deviation from the normal distribution is found, but already at $R = 100$ it is obvious that we have to reject $E(\hat{\beta}_2) = \beta_2$, because the corresponding test statistic is $10 \times (0.409 - 0.5)/0.197 = -4.62$. Although it is doubtful that at $R = 100$ we can already fully rely on the CLT approximation for using critical values from the standard normal distribution, it seems that this deviation from 0.5 of well over 4 times the standard error is very strong evidence of negative bias.

For $R = 10{,}000$ we do find in Figure 2.5 that the distributions of both $\hat{\beta}_1$ and $\hat{\beta}_2$ deviate significantly from the normal. Apparently, in

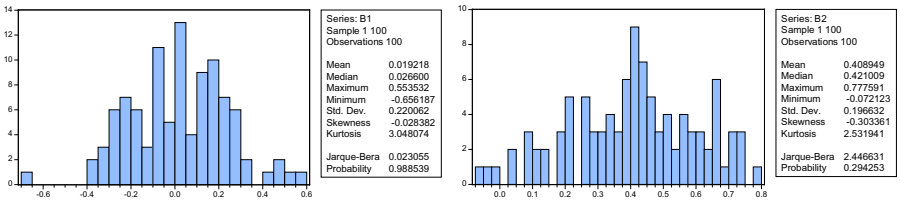

Fig. 2.4 MCS results for program mcs23.prg at $n = 25$, $R = 100$.

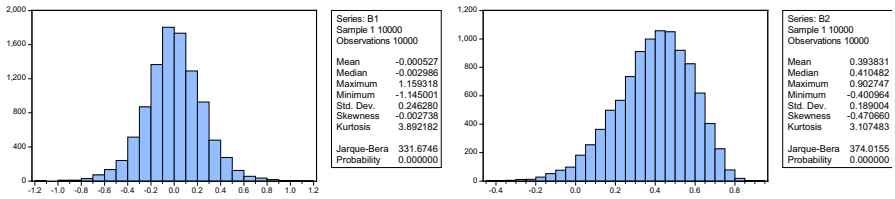

Fig. 2.5 MCS results for program mcs23.prg at $n = 25$, $R = 10{,}000$.

Figure 2.4 the Jarque–Bera tests lacked power[4] due to the small value of R. Although $\hat{\beta}_1$ does not seem notably biased nor skewed, its tails are clearly fat. It is obvious now that $\hat{\beta}_2$ is skew to the left and shows a substantial negative bias, which is with high probability in the range $[-0.1119, -0.1005]$, so about -20% of β_2. For this interval we used again critical value 3, so it has a confidence coefficient of over 99.5% if the normal approximation of the CLT applies[5] indeed for $R = 10{,}000$. However, by program msc21.prg we have demonstrated that for distributions with mild skewness similar to that of B2 even a much smaller Monte Carlo sample size would allow to use the normal approximation for a Studentized sample average, but using a smaller R would of course lead to a wider confidence interval for the bias $E(\hat{\beta}_2) - \beta_2$.

Figure 2.5 shows that the deviations in finite sample of the actual properties of the distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ from those suggested by

[4] This illustrates the established wisdom that one should never simply accept a not rejected null hypothesis. Figure 2.4 did not reject normality of $\hat{\beta}_1$ and $\hat{\beta}_2$ for which we have now established very convincingly in Figure 2.5 that they are actually nonnormal. Note that the actual population distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ do not alter when we change R, while keeping n and β fixed. Hence, the histograms in Figures 2.4 and 2.5 refer to precisely the same nonnormal distributions.

[5] Note that we want the normal approximation to be appropriate not for the skew B2 itself, but for the mean of the R drawings in Series B2.

asymptotic theory are much more moderate for $n = 250$. Although their distributions still deviate significantly from the normal and $\hat{\beta}_2$ is still significantly negatively biased, the absolute magnitude of this bias is not very large now (about -2% of β_2).

In this model, assuming $y_0 \sim N(\bar{y}_0, \sigma_0^2)$, one can derive by higher-order asymptotic expansion methods, see Kiviet and Phillips (2012), that

$$E(\hat{\beta}_2 - \beta_2) = -\frac{1 + 3\beta_2}{n} - \frac{1 - 3\beta_2 + 9\beta_2^2}{n^2(1 - \beta_2)} + \frac{1 + 3\beta_2}{n^2\sigma^2}\left[\sigma_0^2 + \left(\bar{y}_0 - \frac{\beta_1}{1 - \beta_2}\right)^2\right] + O(n^{-3}).$$

For the present version of the program, where β_1, $\bar{y}_0$, and σ_0^2 are all zero, this simplifies to

$$E(\hat{\beta}_2 - \beta_2) = -\frac{1 + 3\beta_2}{n} - \frac{1 - 3\beta_2 + 9\beta_2^2}{n^2(1 - \beta_2)} + O(n^{-3}),$$

and for $\beta_2 = 0.5$ this yields

$$E(\hat{\beta}_2 - \beta_2) = -\frac{2.5}{n} - \frac{3.5}{n^2} + O(n^{-3}).$$

Hence, a first-order asymptotic approximation to the bias of $\hat{\beta}_2$ for $\beta_2 = 0.5$ is -0.1 for $n = 25$ and -0.01 for $n = 250$. Note that these analytic approximations are extremely accurate, as is demonstrated by the simulation findings.

Finally, we will have a look at the accuracy obtained for the various moments when choosing $R = 10{,}000$. Results are collected in Table 2.2

In the columns labeled MCS the actual moment estimates are presented; these, and the calculated values for $\overrightarrow{\tau}$ and $\overrightarrow{\epsilon}$ have all been obtained from Figures 2.5 and 2.6 and from the formulas in (2.31) through (2.35). Note that the accuracy of $\overrightarrow{E}(\hat{\beta}_2)$ is much better for the larger sample size n, because increasing n leads to a sharp decrease in $\mathrm{Var}(\hat{\beta}_2)$. However, regarding the (square roots of the) higher-order moments there is not a similar improvement in the relative precision. When aiming at $\epsilon = 0.01$ then regarding $\overrightarrow{E}(\hat{\beta}_2)$ one could do with fewer replications for n large. But, irrespective of the value of n the RMSE

Table 2.2. Accuracy of MCS moment estimates obtained by msc23.prg for $R = 10{,}000, \beta_1 = 0, \beta_2 = 0.5, y_0 = 0$.

	$n = 25$			$n = 250$		
moment	MCS	$\overrightarrow{\tau}$	$\overrightarrow{\epsilon}$	MCS	$\overrightarrow{\tau}$	$\overrightarrow{\epsilon}$
$\overrightarrow{E}(\hat{\beta}_2)$	0.3938	0.0057	0.0113	0.4894	0.0016	0.0033
$\overrightarrow{\mathrm{SD}}(\hat{\beta}_2)$	0.1890	0.0040	0.0211	0.0545	0.0012	0.0211
$\overrightarrow{\mathrm{Var}}(\hat{\beta}_2)$	0.0357	0.0015	0.0421	0.0030	0.0001	0.0422
$\overrightarrow{\mathrm{MSE}}(\hat{\beta}_2)$	0.0470	0.0023	0.0480	0.0031	0.0001	0.0438
$\overrightarrow{\mathrm{RMSE}}(\hat{\beta}_2)$	0.2168	0.0052	0.0240	0.0558	0.0012	0.0218

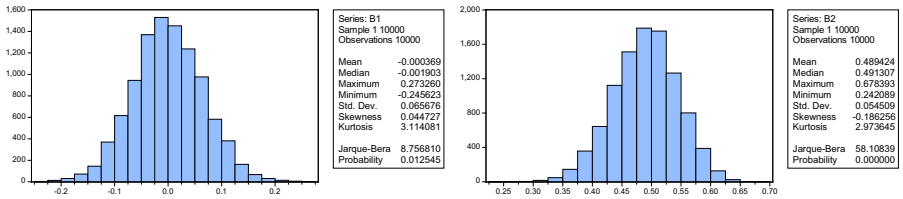

Fig. 2.6 MCS results for program mcs23.prg at $n = 250$, $R = 10{,}000$.

results would need roughly 4 times as many replications and the MSE results even at least 16 times as many. Below we will show, however, that one does not often need very precise MCS estimates of (R)MSE values as such, but only for the differences between the (R)MSE's of alternative econometric estimation techniques, which may require fewer replications than suggested here.

Exercises

1. Prove that $\overrightarrow{Var}_R(q_n)$ of (2.8) and $\overrightarrow{\mathrm{Var}}_R(\overrightarrow{E}_R(q_n))$ of (2.9) are unbiased MCS estimators.
2. Explain why (2.13) is approximately Student $(R-1)$ distributed.
3. Suppose one runs a simulation study on $\overrightarrow{E}_R(\hat{\beta}_j)$ for $R = 1{,}000$ twice, using different independent random numbers (by changing the value of rndseed). Then the successive outcomes of $\overrightarrow{E}_R(\hat{\beta}_j)$ will vary. Argue why they will vary usually by (much) less than 10%.
4. Adapt program mcs21.prg such that x=(@rchisq(1)-1)/@sqrt(2) and w=p*(x+10)-(1-p)*(x+10). Examine, by

taking $R = 1$, the nature of the distributions of these x and w. Note that the $\chi^2(1)$ distribution is even more skew than $\chi^2(2)$. For $v \sim N(0,1)$ one can derive $E(v^r) = [r!/(r/2)!]/2^{r/2}$ for even r, whereas $E(v^r) = 0$ for r odd. From this it follows that $E(x_i) = 0$, $E(x_i^2) = 1$, $E(x_i^3) = 2\sqrt{2} = 2.828$, and $E(x_i^4) = 15$. Why is it that you find values very close to, but nevertheless different from these in your histogram of variable x? What MCS sample size seems required in order for averages of IID variables distributed as x or w to be (almost) normally distributed?

5. Rerun program mcs22.prg (choose Quiet execution mode, because this is much faster). Make the window of workfile mcs22.wf1 active, and double click eq1. Check the values of b1, b2, s, and StdErrb2. Next double click simres and make sure that you recognize the position of these values of b1, b2, s, and StdErrb2 again. Next make the window of workfile mcs21sim.wf1 active. Double click b2 and again note the position of the Rth value of b2 in the spreadsheet. Next rerun the program with a different value of rndseed. Note that the Rth replication is different now, and also the results in the histograms of the variables in mcs21sim.wf1. Are the overall conclusions drawn on the basis of the simulations qualitatively similar? You could also increase/decrease the value of R and examine the effects.
6. Calculate the elements of $\sigma^2(X'X)^{-1}$ for the model of program mcs22.prg. Note that here $X'X = \frac{n}{20}X_0'X_0$, where X_0 contains the first 20 observations of X. The elements of $X_0'X_0$ can be obtained by using $\sum_{t=1}^{T} 1 = T$, $\sum_{t=1}^{T} t = \frac{1}{2}T(T+1)$, and $\sum_{t=1}^{T} t^2 = \frac{1}{6}T(T+1)(2T+1)$.
7. Test the hypothesis $E(s) = \sigma$ using the results from program mcs22.prg. We know that in the present model $E(s^2) = \sigma^2$. The nonlinearity of the transformation $s = \sqrt{s^2}$ implies that s will not be unbiased. Note, however, that this will not imply that we will always reject the hypothesis $E(s) = \sigma$, because the bias of s may be very small and therefore the power of the test may be poor unless $R \to \infty$.

8. Adapt program mcs22.prg such that a final line is added in which you create another variable, namely genr scaledb2=@sqrt(!n)*(b2-!beta2). Choose your own value of rndseed, take !R=10,000, and increase or decrease the value of !sigma. Next run the program for two different values of n, namely !n=40 and a much larger value, for instance !n=400. Compare the two histograms of the distribution of MCS variable scaledb2 for the two different n values and draw the conclusions that seem most relevant to you.
9. Explain that because in MCS one can increase R without much costs, one should not stick to the habitual significance level of 5%. Hence, argue why by taking a very large R one can reduce the risks of both type I and type II errors jointly.
10. Consider the values of Probability (p-value) mentioned for the Jarque–Bera test in EViews histograms. Clarify in words/formulas what the actual values of this probability represent. Explain clearly what you learn from the actual p-values mentioned in Figures 2.5 and 2.6.
11. Derive the values of τ and ϵ regarding the bias assessment of the intercept in the program mcs22.prg for $R = 10{,}000$. Verify whether the actual MCS errors are smaller.
12. Write down the formula for $\hat{\beta}_2$ in the AR(1) model and argue why this expression will not have a normal distribution and not an expectation equal to β_2.
13. Rerun program mcs23.prg, but choose $\beta_2 = 0.95$ and experiment with the choice of y_0 and verify the accuracy of the higher-order asymptotic approximation to the bias. First choose $R = 1{,}000$ and analyze the accuracy of your assessment of the bias. What should R be to guarantee that the first moment of $\hat{\beta}_2$ is obtained with the corresponding ϵ not exceeding 0.02?
14. From the numerical results of program mcs22 and the formulas of Section 2.3 one can find $\overrightarrow{\text{SD}}[\overrightarrow{\text{Var}}(\hat{\beta}_2)] = \overrightarrow{\text{Var}}(\hat{\beta}_2)[(\overrightarrow{\kappa}(\hat{\beta}_2) - 1)/R]^{1/2} = (0.033565)^2 \times (0.0002014)^{1/2} = 0.000016$. Does this make the difference between $\overrightarrow{\text{Var}}(\hat{\beta}_2)$ and $Var(\hat{\beta}_2)$ plausible?

3

Monte Carlo Assessment of Probabilities and Quantiles

The MCS analysis of an expectation or of higher-order (centered) moments presupposes that these moments of q_n do exist. Irrespective of this being the case, the Monte Carlo sample $(q_n^{(1)}, \ldots, q_n^{(R)})$ can always be used fruitfully to estimate the CDF $F_{q_n}(q) \equiv \Pr(q_n \leq q)$ of q_n at any relevant real value q, as we shall show. And, provided that $F_{q_n}(q)$ is well behaved (continuous), MCS also allows to assess its inverse, the quantile function $Q_{q_n}(p) \equiv F_{q_n}^{-1}(p)$ for $0 \leq p \leq 1$. So, even if the expectation does not exist, in principle one can obtain a MCS estimate of the median $Q_{q_n}(0.5)$, also known as the second quartile. And when the variance does not exist one can express the spread of the distribution by for instance the interquartile range (IQR), which is $Q_{q_n}(0.75) - Q_{q_n}(0.25)$, or the third minus the first quartile. Before we get to that, we first focus on estimating probabilities $\Pr(q_n \leq q)$ by MCS, not just for the case where q_n is an estimator and $\Pr(q_n \leq q)$ its CDF, but especially for the case where q_n is a test statistic. Then $\Pr(q_n \leq q)$ or $\Pr(q_n \geq q) = 1 - \Pr(q_n < q)$ may represent the (one-sided) rejection probability at critical value q, or a so-called p-value when q is actually an observed empirical realization of the test statistic.

3.1 MCS Estimation of a Probability

The CDF $F_{q_n}(q)$ of statistic q_n can for any real value q be estimated by MCS by establishing its empirical counterpart, to be denoted as $\overrightarrow{F}_{R,q_n}(q)$, from the simulated sample $(q_n^{(1)},\ldots,q_n^{(R)})$. From the R independent drawings from the distribution of q_n the MCS estimate of the CDF is

$$\overrightarrow{F}_{R,q_n}(q) \equiv \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}(q_n^{(r)} \le q), \tag{3.1}$$

where $\mathbb{I}(\cdot)$ is the 0–1 indicator function, which is 1 when $q_n^{(r)} \le q$ is true and 0 otherwise. Note that for any q the R random variables $d^{(r)} \equiv \mathbb{I}(q_n^{(r)} \le q)$ constitute an IID sample of Bernoulli drawings with

$$\begin{aligned} E(d^{(r)}) &= 1 \times \Pr(q_n^{(r)} \le q) + 0 \times \Pr(q_n^{(r)} > q) = F_{q_n}(q) \\ \mathrm{Var}(d^{(r)}) &= 1^2 \times \Pr(q_n^{(r)} \le q) + 0^2 \times \Pr(q_n^{(r)} > q) - [E(d^{(r)})]^2 \\ &= F_{q_n}(q)[1 - F_{q_n}(q)]. \end{aligned}$$

Thus, $\overrightarrow{F}_{R,q_n}(q) = R^{-1}\sum_{r=1}^{R} d^{(r)}$ is an unbiased estimator of $F_{q_n}(q)$, because

$$E[\overrightarrow{F}_{R,q_n}(q)] = \frac{1}{R}\sum_{r=1}^{R}E(d^{(r)}) = F_{q_n}(q) \tag{3.2}$$

with

$$\begin{aligned} \mathrm{Var}[\overrightarrow{F}_{R,q_n}(q)] &= \mathrm{Var}\left(\frac{1}{R}\sum_{r=1}^{R} d^{(r)}\right) = \frac{1}{R^2}\sum_{r=1}^{R}\mathrm{Var}(d^{(r)}) \\ &= \frac{1}{R}F_{q_n}(q)[1 - F_{q_n}(q)], \end{aligned} \tag{3.3}$$

where the second equality follows from the independence of the random variables $d^{(r)}$. Hence, also for estimating probabilities by MCS the variance of the unbiased MCS estimator is $O(R^{-1})$ so that by increasing R we can achieve any degree of accuracy that we want. Below, we will no longer supply $\overrightarrow{F}$ with subindex R.

The LLN yields $\mathrm{plim}_{R\to\infty}\overrightarrow{F}_{q_n}(q) = F_{q_n}(q)$. That the empirical CDF converges in every q to the true population CDF is known as the

Fundamental Theorem of Statistics. By the CLT we find

$$\sqrt{R}(\overrightarrow{F}_{q_n}(q) - F_{q_n}(q)) \underset{R\to\infty}{\overset{d}{\to}} N(0, F_{q_n}(q)[1 - F_{q_n}(q)]), \tag{3.4}$$

from which it follows that

$$\frac{\overrightarrow{F}_{q_n}(q) - F_{q_n}(q)}{\{\overrightarrow{F}_{q_n}(q)[1 - \overrightarrow{F}_{q_n}(q)]/R\}^{1/2}} \overset{a}{\sim} N(0,1). \tag{3.5}$$

This can be used to construct a confidence interval for the probability $F_{q_n}(q)$ and to derive the required value of R in order to achieve a particular chosen limited width of this confidence interval.

3.1.1 Assessing Rejection Probabilities

Let the $p \times 1$ vector θ again represent the full vector of parameters determining the DGP that is simulated. Now if q_n is a scalar test statistic for a particular null-hypothesis, say $g(\theta) = 0$, whereas it has been derived that asymptotically (for $n \to \infty$) test statistic q_n has a particular null-distribution which has at the nominal significance level α a right-hand tail critical value c_R^α, then we can use Monte Carlo experiments to verify what the actual type I error probability of the test will be for finite n. In the experiments θ should be chosen such that the null is true. When $g(\theta)$ is a vector function of dimension less than p, which it usually is, this implies that under the null θ is in some unbounded subspace of $\mathbb{R}^p$. Unless invariance properties of the test statistic over this subspace are available, establishing the actual size of the test requires that this whole subspace is scanned in the Monte Carlo study. The complications involved will be clarified further in the next section.

For a particular θ such that $g(\theta) = 0$, a confidence set (again with high confidence coefficient of at least 99.5%, provided R is so large that the approximation by the CLT is accurate) for the actual type I error probability is given now by the endpoints

$$1 - \overrightarrow{F}_{q_n}(c_R^\alpha) \pm 3[\overrightarrow{F}_{q_n}(c_R^\alpha)(1 - \overrightarrow{F}_{q_n}(c_R^\alpha))/R]^{1/2}. \tag{3.6}$$

A corresponding test of the hypothesis $1 - F_{q_n}(c_R^\alpha) = \alpha$ can be performed by the test statistics

$$\frac{1 - \overrightarrow{F}_{q_n}(c_R^\alpha) - \alpha}{[\overrightarrow{F}_{q_n}(c_R^\alpha)(1 - \overrightarrow{F}_{q_n}(c_R^\alpha))/R]^{1/2}} \quad \text{or} \quad \frac{1 - \overrightarrow{F}_{q_n}(c_R^\alpha) - \alpha}{[\alpha(1-\alpha)/R]^{1/2}}, \tag{3.7}$$

which should be compared with a critical value from the standard normal, mostly at a significance level much smaller than the α as used in c_R^α. When the critical region of the test statistic is to the left of c_L^α these test statistics have of course $\overrightarrow{F}_{q_n}(c_L^\alpha) - \alpha$ in their numerators.

The formulas for a test with a right-hand tail critical value become slightly simpler by defining the survivor function $S_{q_n}(q) \equiv \Pr(q_n > q) = 1 - F_{q_n}(q)$, which is 1 when $q_n > q$ and 0 otherwise, and then substituting the empirical survivor function $\overrightarrow{S}_{q_n}(q) \equiv R^{-1}\sum_{r=1}^{R} \mathbb{I}(q_n^{(r)} > q) = 1 - \overrightarrow{F}_{q_n}(q)$. After generating in EViews the rejection indicator variable $d_n^{*(r)} \equiv 1 - d_n^{(r)} = \mathbb{I}(q_n^{(r)} > q)$ one easily obtains the descriptive statistics

$$\text{Mean} = \overrightarrow{S}_{q_n}(q) = 1 - \overrightarrow{F}_{q_n}(q) \tag{3.8}$$

and

$$\text{Std. Dev.} \approx \{\overrightarrow{F}_{q_n}(q)[1 - \overrightarrow{F}_{q_n}(q)]\}^{1/2} = \{\overrightarrow{S}_{q_n}(q)[1 - \overrightarrow{S}_{q_n}(q)]\}^{1/2}. \tag{3.9}$$

The latter approximation holds because for large R we have $R/(R-1) \approx 1$ and for $\bar{d} \equiv R^{-1}\sum_{r=1}^{R} d^{(r)} = \overrightarrow{F}_{q_n}(q)$ we have $(R-1) \times (\text{Std. Dev.})^2 = \sum_{r=1}^{R}(d^{(r)} - \bar{d})^2 = \sum_{r=1}^{R} d^{(r)} - R\bar{d}^2 = R\bar{d}(1-\bar{d}) = R\overrightarrow{F}_{q_n}(q)[1 - \overrightarrow{F}_{q_n}(q)]$, since $d^{(r)}d^{(r)} = d^{(r)}$ for a 0–1 variable. By dividing this value of Std. Dev. for variable d_n or d_n^* by $\sqrt{R}$ one obtains

$$\overrightarrow{\text{SD}}(\overrightarrow{F}_{q_n}(q)) = \overrightarrow{\text{SD}}(\overrightarrow{S}_{q_n}(q)) = \{\overrightarrow{F}_{q_n}(q)[1 - \overrightarrow{F}_{q_n}(q)]/R\}^{1/2}. \tag{3.10}$$

A similar approach as the one given above can be followed to assess the rejection probability when the null hypothesis is not true, i.e., $g(\theta) \neq 0$. As a rule such probabilities vary with θ, so to get an impression of the dependence of the rejection probability when the null hypothesis is not true may require extensive calculations in order to cover a relevant grid of values in the usually unbounded parameter space under the alternative hypothesis. These rejection probability estimates should not directly be interpreted as power unless it has been established that the test is not affected by so-called size distortions. In that case, however, an impression of power can be obtained after a size correction, for which quantiles have to be assessed first. Size correction and power comparisons are discussed in detail in Section 4.

3.1.2 Choosing the MCS Sample Size

Again, note that by choosing R sufficiently large the width of interval (3.6) and the type II error probability of the tests (3.7) can be made arbitrarily small. Suppose that we want to use $\overrightarrow{F}_{q_n}(c_R^\alpha)$ to assess (again with a high confidence level of 99.5%) whether the actual type I error probability of a test using critical value c_R^α deviates more than $100\epsilon\%$ from the nominal significance level α (which could be 5% or any other arbitrary value). Then the number of replications R has to be such that

$$3[\alpha(1-\alpha)/R]^{1/2} \leq \epsilon\alpha \quad \text{or} \quad R \approx \frac{1-\alpha}{\alpha}\frac{9}{\epsilon^2}. \tag{3.11}$$

So, for $\epsilon = 0.01$ and $\alpha = 0.05$ we need $R \approx 2 \times 10^6$, whereas for $R =$ 10,000 one can establish type I error probabilities of about 5% by an error margin of only about 13%. Thus, when $R = 10{,}000$, an estimated type I error probability within the interval $[0.043, 0.057]$ corroborates that the actual type I error probability could be 5%. From this we conclude that, when employing the same precision measure ϵ, in general one needs a much larger MCS sample size R for analyzing correspondence between actual and nominal type I error probabilities than for verifying whether an estimator seems biased. However, when determining R in case of an analysis of bias there is an extra complication, because the outcome depends on the unknown $\mathrm{Var}(\overrightarrow{E}(q_n)) = \mathrm{Var}(q_n)/R$, whereas the variance of $\overrightarrow{F}_{q_n}(c_R^\alpha)$ equals $F_{q_n}(c_R^\alpha)(1 - F_{q_n}(c_R^\alpha))/R$, which in case of little or no size problems is close to $\alpha(1-\alpha)/R$.

When determining R, though, it should not just be the variance of $\overrightarrow{F}_{q_n}(c_R^\alpha)$ that should bother us. The magnitude of R is also relevant for whether or not we can rely on the CLT when using critical values from the standard normal distribution for constructing confidence intervals (3 for a confidence coefficient of at least 99.5%). Hence, R does not just determine the width of MCS confidence intervals as such, but also their actual confidence coefficient. Before we employ the above theory on estimating probabilities we shall first focus on what values of R are required in order to rely safely on the normal approximation of the CLT for statistics similar to $\overrightarrow{F}_{q_n}(q)$.

3.1.3 CLT and the Sample Average of Binary Variables

To examine the speed of convergence of a MCS probability estimate to normality, we use the following program:

```
'mcs31.prg:  another illustration of the CLT
rndseed 9876543210
!R=10000
workfile f:\MCS\mcs31.wf1 u 1 !R
!metaR=100000
matrix (!metaR,4) simres
for !metarep=1 to !metaR
   genr u=rnd
   genr da=u<0.5
   genr db=u<0.1
   genr dc=u<0.01
   genr dd=u<0.001
   simres(!metarep,1)=@mean(da)
   simres(!metarep,2)=@mean(db)
   simres(!metarep,3)=@mean(dc)
   simres(!metarep,4)=@mean(dd)
next
simres.write f:\MCS\mcs31sim.txt
workfile f:\MCS\mcs31sim.wf1 u 1 !metaR
read f:\MCS\mcs31sim.txt pa pb pc pd
genr scaledpa=@sqrt(!R)*(pa-0.5)/0.5
genr scaledpb=@sqrt(!R)*(pb-0.1)/@sqrt(0.09)
genr scaledpc=@sqrt(!R)*(pc-0.01)/@sqrt(0.0099)
genr scaledpd=@sqrt(!R)*(pd-0.001)/@sqrt(0.000999)
```

Note that we use MCS here to estimate probabilities from a sample of size R which have true value $p_a = 0.5$, $p_b = 0.1$, $p_c = 0.01$, and $p_d = 0.001$, respectively. We verify how well this works in a meta MCS study with sample size $metaR = 100{,}000$ by examining $metaR$ realizations of the estimation errors $\overrightarrow{p}_i - p_i$, for $i = a, b, c, d$. We find the following histograms from the meta MCS study for $(\overrightarrow{p}_i - p_i)/[p_i(1 - p_i)/R]^{1/2}$. Taking $R = 100$ the results are as in Figure 3.1.

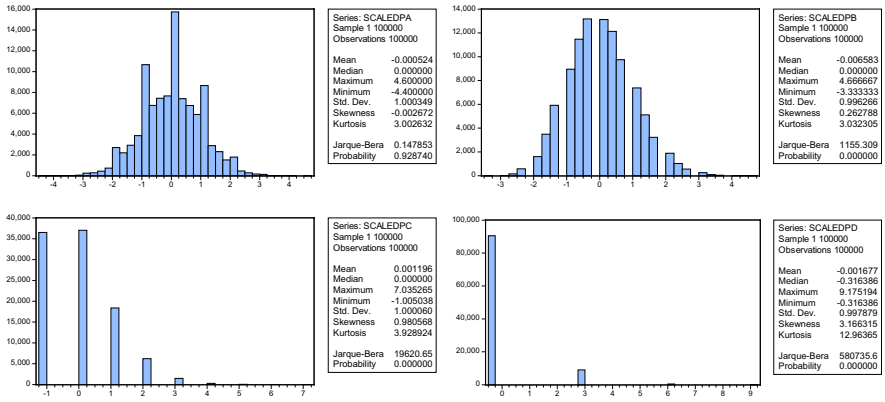

Fig. 3.1 Estimating probabilities from $R = 100$ drawings.

Figure 3.1 shows that for $R = 100$ the MCS estimator of a probability of 0.5 has its first four moments reasonably close to those of a standard normal distribution, but is nevertheless clearly different. The smaller the estimated probability gets, the worse the normal approximation is, and the more apparent it becomes that we should actually start off from the true distribution of $R\overrightarrow{p}_i = \sum_{r=1}^{R} d^{(r)}$, which is binomial. The binomial distribution is nonsymmetric for $p \neq 0.5$ and the Normal approximation is known to be deficient when p or 1–p is very small, even when R is substantial. From Figure 3.1 we observe, however, that relatively few of the depicted scaled errors are in fact larger in absolute value than 3.

From Figure 3.2 we see that when we use 1,000 replications it seems acceptable to use critical values at very small significance levels from the normal approximation when we are estimating a probability in the range 0.1 through 0.5. However, estimating probabilities in the range 0.001 through 0.01 requires certainly much more than 1,000 replications, not just for approximating the estimator by the normal distribution, but also because otherwise the standard error of the estimate is too large. Even at sample size 10,000 we have $\sqrt{0.999 * 0.001/10{,}000} = 0.0003$, hence 30%, whereas for a probability of 0.5 it is $\sqrt{0.5^2/10{,}000} = 0.005$, which is only 1%.

Figure 3.3 gives results from program mcs31.prg for $R = 10{,}000$. From this we learn that for estimating probabilities not much smaller

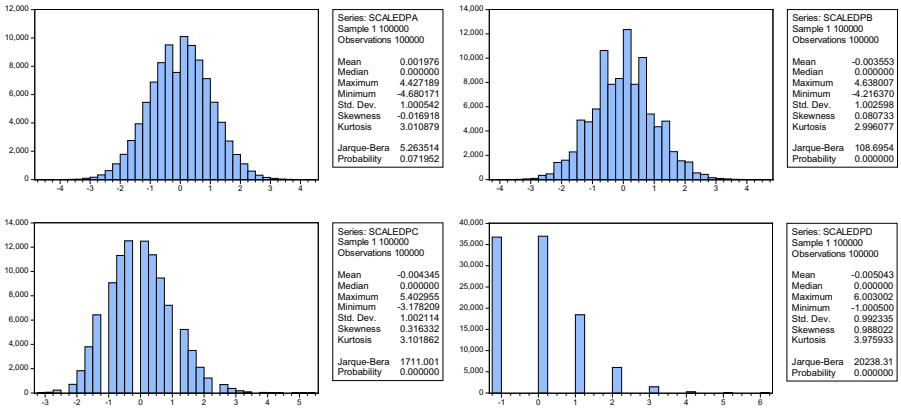

Fig. 3.2 Estimating probabilities from $R = 1{,}000$ drawings.

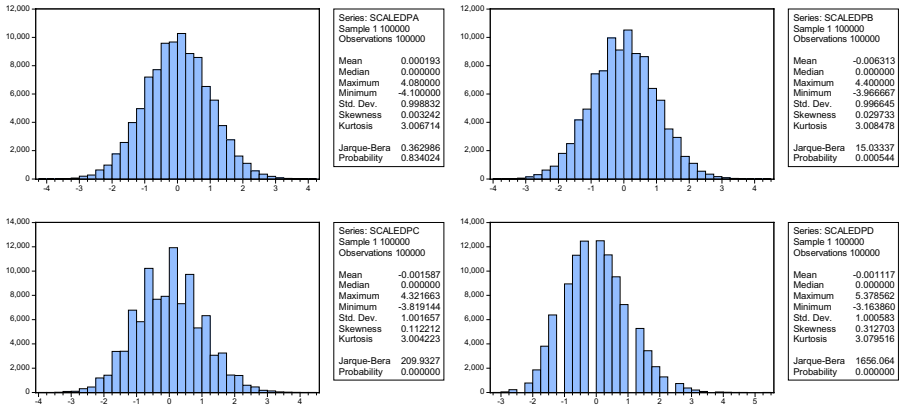

Fig. 3.3 Estimating probabilities from $R = 10{,}000$ drawings.

than 0.01 (or larger than 0.99) using the normal approximation with critical value 3 will yield confidence intervals with certainly a very high confidence coefficient.

The reassurance just obtained regarding the extremely high confidence that can be associated to using a critical value c, with $c \geq 2.8$, inspires the following simple rule of thumb. When estimating a probability p by MCS one can be pretty sure that the first $d \geq 1$ decimals are correct, if R satisfies

$$c\sqrt{p(1-p)/R} \leq a(0.1)^d \quad \text{or} \quad R \geq (c/a)^2 p(1-p)10^{2d},$$

where a safe choice of a is $a \leq 0.1$. Because $\max p(1-p) = 0.25$, this yields

$$R \approx 2 \times 10^{2(d+1)}. \tag{3.12}$$

This again underscores that, unless methods are exploited that enhance their efficiency (see Sections 5.2.1 and 5.4), for reasonably accurate estimates of probabilities one might need at least a million replications.

3.1.4 More on MCS Sample Size when Estimating Type I Error Probabilities

MCS is often used to verify whether the actual significance level of a test is close to the chosen nominal value, where popular values of the latter are 5%, 2.5% or sometimes 1%. When using 10,000 replications probabilities of 0.05, 0.025, and 0.01 are estimated with standard errors of 0.002, 0.0016, and 0.001, respectively. Hence, especially for a probability of 0.01 the accuracy is limited. However, for most practical purposes $R = 10,000$ seems fine, because when we use a test in econometrics at nominal significance level of 5% and the actual risk of type I errors is in fact 4% or 6% this is usually of minor concern. Though, in most cases the situation is problematic when either the actual significance level is larger than $\frac{4}{3}\alpha$ (much higher risk of type I errors than aimed at) or smaller than $\frac{2}{3}\alpha$ (much higher risk of type II errors than intended). That such differences, if they exist, will be noted with much higher chance when $R = 10,000$ than when one uses $R = 1{,}000$ becomes clear from the probabilities collected in Table 3.1. These are exact, because they have not been obtained by simulation or approximation, but from calculations fully respecting the Binomial nature of the underlying probability distribution.

From Table 3.1 it can be seen that when trying to assess actual type I error probabilities in the range 2.5% through 5% within a particular limited margin it is risky to use 1,000 or fewer replications, whereas choosing 10,000 is safe for most practical purposes, whereas 5,000 seems pretty reasonable too. When one is interested in probabilities close to 1% then one may need more than 10,000 replications in order to realize reasonable accuracy.

Table 3.1. Probabilities $\Pr\{\overrightarrow{p} \in \mathcal{I}\}$ for a few particular $p, \mathcal{I}, R$.

p	$\mathcal{I}$	$R = 1{,}000$	$R = 5{,}000$	$R = 10{,}000$
0.10	[0.09, 0.11]	0.7318	0.9827	0.9991
	[0.08, 0.12]	0.9894	1.000	1.000
	[0.07, 0.13]	0.9986	1.000	1.000
0.05	[0.045, 0.055]	0.5752	0.9022	0.9795
	[0.04, 0.06]	0.8731	0.9989	1.000
	[0.03, 0.07]	0.9970	1.000	1.000
0.025	[0.0225, 0.0275]	0.3871	0.7427	0.8978
	[0.02, 0.03]	0.7358	0.9792	0.9988
	[0.015, 0.035]	0.9680	1.000	1.000
0.01	[0.009, 0.011]	0.3657	0.5657	0.7089
	[0.008, 0.012]	0.5735	0.8652	0.9609
	[0.005, 0.015]	0.9234	0.9996	1.000

3.2 Accuracy of Inference in Nonnormal Linear Regression

We will now set out again to use MCS to a situation where for some aspects of the distribution of OLS based regression inference, no clear-cut analytical finite sample results are available. We will look at the classic linear regression model with nonnormal disturbances. Hence, the Gauss–Markov conditions are satisfied such that OLS is the BLUE (best linear unbiased estimator), but in finite samples the coefficient estimator is not normally distributed and thus even under the null hypothesis t-statistics will not follow Student's t distribution, so that the actual significance level of a test may deviate from the nominal level. In practice this means that one will use (first-order) asymptotic approximations, because under mild regularity even when the disturbances are nonnormal the coefficient estimates will be asymptotically normal, and the null distribution of t-tests will be asymptotically standard normal. We use the following program:

```
'mcs32.prg:  MCS of a classic regression model with nonnormal disturbances
!n=20
workfile f:\MCS\mcs32.wf1 u 1 !n
genr i=@trend(0)
genr x2=i-20*@floor((i-1)/20)
genr x3=2+0.25*x2^2
!beta1=5  'intercept
!beta2=0.8  'slope
!sigma=1.5
rndseed 9876543210
```

```
!R=2500
matrix (!R,6) simres
for !rep=1 to !R
   genr u=!sigma*(@rchisq(1)-1)/@sqrt(2)
   genr y=!beta1+!beta2*x2+u
   equation eq1.ls y c x2 x3
   simres(!rep,1)=eq1.@coefs(3)
   simres(!rep,2)=eq1.@stderrs(3)
   simres(!rep,3)=eq1.@coefs(2)
   simres(!rep,4)=eq1.@stderrs(2)
   equation eq2.ls y c x2
   simres(!rep,5)=eq2.@coefs(2)
   simres(!rep,6)=eq2.@stderrs(2)
next
simres.write f:\MCS\mcs32sim.txt
workfile f:\MCS\mcs32sim.wf1 u 1 !R
read f:\MCS\mcs32sim.txt b3 seb3 b2 seb2 b2e seb2e
genr tb3=b3/seb3
genr rejectb3LR= abs(tb3)>@qtdist(0.975,!n-3)
genr rejectb3L= tb3<@qtdist(0.05,!n-3)
genr rejectb3R= tb3>@qtdist(0.95,!n-3)
genr tb2=(b2-!beta2)/seb2
genr rejectb2LR= abs(tb2)>@qtdist(0.975,!n-3)
genr rejectb2L= tb2<@qtdist(0.025,!n-3)
genr rejectb2R= tb2>@qtdist(0.975,!n-3)
genr tb2e=(b2e-!beta2)/seb2e
genr rejectb2eLR= abs(tb2e)>@qtdist(0.95,!n-2)
genr rejectb2eL= tb2e<@qtdist(0.05,!n-2)
genr rejectb2eR= tb2e>@qtdist(0.95,!n-2)
```

3.2.1 Establishing Size Problems

The disturbances have been generated by rescaling drawings from the $\chi^2(1)$ distribution, such that $u_t \sim IID(0, \sigma^2)$ but extremely skew. Note that the DGP is such that x3 is a redundant variable (has coefficient zero), so its t-ratio should ideally have a distribution such that we reject $H_0 : \beta_3 = 0$ with a probability close to the value that we find appropriate and therefore choose as nominal significance level. Note that it can be proved that estimator b2e (obtained in the regression where x3 has been omitted) is more efficient than b2 (because x2 and x3 are not orthogonal). We first examine the case $n = 20$. Figure 3.4 provides histograms of the coefficient estimators and test statistics.

These results corroborate the unbiasedness of the OLS estimators in this model, because $H_0 : E(\hat{\beta}_j) = \beta_j$ is not rejected. Also note that

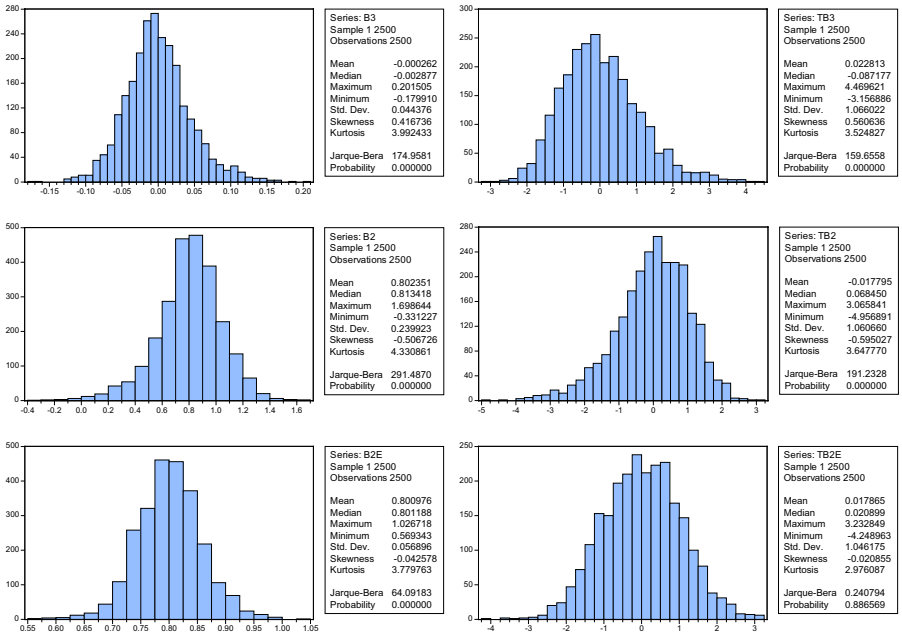

Fig. 3.4 Effects of skew disturbances on estimators and test statistics, $n = 20$.

the 2500 observations of series B2E have substantially smaller standard error than those of B2. It is also obvious that the coefficient estimators are nonnormal and especially the t-statistics of B2 and B3 are found to have a skew distribution (but with an expectation close to zero).

Regarding the rejection indicator variables EViews provides the descriptive statistics mean and Std. Dev. as indicated in Table 3.1. These mean values establish estimates of rejection probabilities, which have standard error (Std. Err.) equal to the given Std. Dev. values divided by $\sqrt{R} = 50$. The table also gives the nominal significance levels of the various tests (which follow from the critical values used in the program).

Although the severe nonnormality of the disturbances, and the very small econometric sample size of $n = 20$, clearly affect the actual null distributions so that they deviate from Student's t, nevertheless their tail-probabilities are such that no serious distortions are found for the type I error probabilities. Apparently, a sample as small as 20 already

Table 3.2. Results from program msc23.prg for $n = 20$, $R = 2{,}500$.

	B2EL	B2ELR	B2ER	B2L	B2LR	B2R	B3L	B3LR	B3R
Mean	0.0432	0.0904	0.0472	0.0392	0.0460	0.0068	0.0284	0.0472	0.0676
Std. Dev.	0.2033	0.2868	0.2121	0.1941	0.2095	0.0822	0.1661	0.2121	0.2511
Std. Err.	0.0041	0.0057	0.0042	0.0039	0.0042	0.0016	0.0033	0.0042	0.0050
Nom. Sig. Lev.	0.0500	0.1000	0.0500	0.0250	0.0500	0.0250	0.0500	0.0500	0.0500

Table 3.3. Type I error probabilities for different n and $R = 1{,}000{,}000$.

		B2EL	B2ELR	B2ER	B2L	B2LR	B2R	B3L	B3LR	B3R
n	nominal	0.0500	0.1000	0.0500	0.0250	0.0500	0.0250	0.0500	0.0500	0.0500
20		0.0486	0.0974	0.0487	0.0376	0.0470	0.0095	0.0312	0.0471	0.0654
200		0.0500	0.1003	0.0503	0.0288	0.0494	0.0206	0.0460	0.0492	0.0536

allows to rely on the blessings of the CLT in the standard analysis of classic regression models. Depending on how accurate one requires asymptotic approximations to be, a MCS sample size of only 2,500 allows already to come to this reassuring conclusion.

We expect the situation to be even better when we increase n. In Table 3.3 we examine both $n = 20$ and $n = 200$ for a very large R in order to improve the accuracy.[1]

We find that at $n = 200$ most tests behave properly, although in the overspecified model (including the redundant regressor x3) the one tailed tests are still a little affected by the skewness of the distribution of the test statistics. Over and underrejection in the two separate tails more or less compensate in the two-tailed tests. The different signs of the skewness of the estimators for β_2 and β_3 lead to opposite behavior in the tails of their respective tests. Note that some of the results for $n = 20$ in Table 3.3 differ substantially from those obtained from the smaller MCS sample, illustrating that 2,500 replications is not very much when estimating an actual significance level. However, using 2,500 replications does suffice to detect really serious deviations from the nominal level.

[1] On a currently rather standard notebook pc obtaining the $R = 10^6$, $n = 200$ simulation results took 15 minutes, whereas the $R = 2{,}500$, $n = 20$ results took 2 seconds.

3.3 MCS Estimation of a Quantile

For $0 < p < 1$ the pth quantile $Q_{q_n}(p)$ of q_n (or its 100th percentile) is given by the solution for $Q_{q_n}(p)$ of

$$\Pr(q_n \leq Q_{q_n}(p)) = F_{q_n}(Q_{q_n}(p)) = p. \tag{3.13}$$

Hence, $Q_{q_n}(0.5)$ is the median of statistic q_n, and the interquartile range is $Q_{q_n}(0.75) - Q_{q_n}(0.25)$. And, when q_n represents a test statistic generated from a DGP in which the tested null hypothesis is true, then its $100\alpha\%$ left-hand critical value is $Q_{q_n}(\alpha)$ and its $100\alpha\%$ right-hand critical value is $Q_{q_n}(1-\alpha)$.

MCS estimates of quantiles of the continuous distribution $F_{q_n}(q)$ can be obtained in the following way. The IID Monte Carlo sample $(q_n^{(1)}, \ldots, q_n^{(R)})$ can be sorted in non-decreasing order, yielding $(q_n^{*(1)}, \ldots, q_n^{*(R)})$, such that

$$q_n^{*(1)} \leq q_n^{*(2)} \leq \cdots \leq q_n^{*(R)}. \tag{3.14}$$

From this $q_n^{*(r)}$ series point estimates for any quantiles can be obtained straightforwardly by establishing the sample or empirical quantiles. A simple MCS estimate of $Q_{q_n}(p)$ is

$$\overrightarrow{Q}_{q_n}(p) \equiv q_n^{*\lceil Rp \rceil}, \tag{3.15}$$

where $\lceil z \rceil$ denotes the integer ceiling of the real positive number z, thus $Rp \leq \lceil Rp \rceil < Rp + 1$. To avoid too many indices we do not dress up $\overrightarrow{Q}_{q_n}(p)$ with an R; the arrow should make clear that Monte Carlo estimation from an IID sample of R replications is involved. It can be derived, see David (1981), that

$$E(\overrightarrow{Q}_{q_n}(p)) = Q_{q_n}(p) + O(R^{-1}) \tag{3.16}$$

and

$$\begin{aligned} \mathrm{Var}(\overrightarrow{Q}_{q_n}(p)) &= \frac{p(1-p)}{R[f_{q_n}(Q_{q_n}(p))]^2} + O(R^{-2}), \quad \text{where} \\ f_{q_n}(Q_{q_n}(p)) &= \frac{\partial}{\partial x} F_{q_n}(q)\Big|_{q=Q_{q_n}(p)}. \end{aligned} \tag{3.17}$$

Due to the assumed continuity of the distribution of q_n, the probability that any of the R values (3.14) are equal is zero. Therefore, these R values define $R+1$ mutually exclusive intervals on the real axis, such that

$$F_{q_n}(q_n^{*(r)}) = \Pr(q_n \leq q_n^{*(r)}) = \frac{r}{R+1}. \tag{3.18}$$

The qualities of quantile estimator (3.15) benefit if R and p are such that $(R+1)p$ is an integer number. This is easily demonstrated for the median. When R is odd, then $\frac{R+1}{2}$ is integer, and $\overrightarrow{Q}_{q_n}(0.5) = q_n^{*(\frac{R+1}{2})}$ is an unbiased estimator. However, when R is even, then $\overrightarrow{Q}_{q_n}(0.5) = q_n^{*(R/2}$, whereas a more appropriate quantile estimator for the median would then be $(q_n^{*(R/2)} + q_n^{*(1+R/2)})/2$, and even more subtile modifications are possible.

It can be shown, that both the LLN and the CLT apply to $\overrightarrow{Q}_{q_n}(p)$. Thus $\text{plim}_{R\to\infty} \overrightarrow{Q}_{q_n}(p) = Q_{q_n}(p)$ and

$$\sqrt{R}(\overrightarrow{Q}_{q_n}(p) - Q_{q_n}(p)) \underset{R\to\infty}{\overset{d}{\to}} N\left(0, \frac{p(1-p)}{[f_{q_n}(Q_{q_n}(p))]^2}\right). \tag{3.19}$$

In order to exploit the latter result we have to estimate $f_{q_n}(Q_{q_n}(p))$, i.e., obtain $\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p))$. Note that we have already seen particular MCS (step-function) results $\overrightarrow{f}_{q_n}(q)$ in the form of histograms as produced by EViews. Since

$$f_{q_n}(Q_{q_n}(p)) = \frac{\partial}{\partial q} F_{q_n}(q)\Big|_{q=Q_{q_n}(p)} = \lim_{q\to Q_{q_n}(p)} \frac{F_{q_n}(q) - F_{q_n}(Q_{q_n}(p))}{q - Q_{q_n}(p)},$$

we will employ

$$\overrightarrow{f}_{q_n}(Q_{q_n}(p)) \equiv \frac{F_{q_n}(Q_{q_n}(p+R^{-1})) - F_{q_n}(Q_{q_n}(p))}{Q_{q_n}(p+R^{-1}) - Q_{q_n}(p)}. \tag{3.20}$$

Exploiting (3.15) and (3.18), this suggests the estimator

$$\begin{aligned}\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p)) &\equiv \frac{F_{q_n}(\overrightarrow{Q}_{q_n}(p+R^{-1})) - F_{q_n}(\overrightarrow{Q}_{q_n}(p))}{\overrightarrow{Q}_{q_n}(p+R^{-1}) - \overrightarrow{Q}_{q_n}(p)} \\ &= \frac{F_{q_n}(q_n^{*\lceil Rp+1\rceil}) - F_{q_n}(q_n^{*\lceil Rp\rceil})}{q_n^{*\lceil Rp+1\rceil} - q_n^{*\lceil Rp\rceil}}\end{aligned}$$

$$= \frac{(\lceil Rp+1 \rceil - \lceil Rp \rceil)/(R+1)}{q_n^{*\lceil Rp+1 \rceil} - q_n^{*\lceil Rp \rceil}}$$
$$= \frac{1}{R+1} \frac{1}{q_n^{*\lceil Rp+1 \rceil} - q_n^{*\lceil Rp \rceil}}. \tag{3.21}$$

In the event (with low probability) that the denominator is (very close to) zero one could take

$$\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p)) = \frac{1}{R+1} \frac{2}{q_n^{*\lceil Rp+1 \rceil} - q_n^{*\lceil Rp-1 \rceil}}.$$

Another option is to use a more sophisticated kernel density estimator, or — as we will demonstrate below — use the histogram of q_n or q_n^*.

Then, exploiting

$$\frac{\overrightarrow{Q}_{q_n}(p) - Q_{q_n}(p)}{[p(1-p)/R]^{1/2} / \overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p))} \overset{a}{\sim} N(0,1), \tag{3.22}$$

a confidence interval for $Q_{q_n}(p)$ can be constructed with high confidence coefficient 99.5%, namely

$$[\overrightarrow{Q}_{q_n}(p) - 3\frac{[p(1-p)/R]^{1/2}}{\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p))}, \quad \overrightarrow{Q}_{q_n}(p) + 3\frac{[p(1-p)/R]^{1/2}}{\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p))}]. \tag{3.23}$$

From this it easily follows how one should determine R in order to achieve an interval that contains the true quantile with high probability and has smaller relative error than $100\epsilon\%$. This requires

$$3\frac{[p(1-p)/R]^{1/2}}{\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p))} \leq \epsilon \overrightarrow{Q}_{q_n}(p)$$

or

$$R \geq \frac{9p(1-p)}{\epsilon^2(\overrightarrow{Q}_{q_n}(p) \times \overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p)))^2}. \tag{3.24}$$

Note that, although formula (3.21) for $\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p))$ involves R, we actually have that $f_{q_n}(Q_{q_n}(p))$ does not vary with R. To use (3.24) for choosing R, first initial assessments of $\overrightarrow{Q}_{q_n}(p)$ and of $\overrightarrow{f}_{q_n}(\overrightarrow{Q}_{q_n}(p))$ are required.

The analysis above should again be taken as a warning that MCS results always involve approximations, because however large R

will be chosen, it will always be finite, whereas the quality of the approximations is not just determined by the magnitude of R, but also by properties of the actual phenomena under investigation. Regarding the accuracy of moment estimation we established that the accuracy depends on the magnitude of the higher-order moments. Here we found for probability estimates that their accuracy depends on the actual magnitude of these probabilities, and for quantiles the actual magnitude of the density in the yet unknown quantile is a determining factor of the accuracy. However, from initial MCS results it can be analyzed what value of R should be appropriate in order to realize particular values for absolute tolerance and relative precision.

3.4 Simulation of Estimators That Have No Moments

Choosing R extremely large does not mean that we can always rely on the CLT without further worries, because the CLT only applies to expressions that do have a finite first moment. Not all estimators used in econometrics do have finite moments. Examples are instrumental variable estimators in just identified models and estimators of long run multipliers in dynamic models. Both are in fact ratios of estimators and problems emerge when the expression in the denominator has a density at argument zero that is nonzero. Then the distribution of this ratio has fat tails and its moments cease to exist, simply because their defining integral does not converge.

We will illustrate this for the total multiplier of a dynamic regression model. Consider the ARX(1) model

$$y_i = \beta_1 + \beta_2 y_{i-1} + \beta_3 x_i + u_i,$$

where $u_i \sim IID(0, \sigma^2)$ and the regressor x is exogenous. The model is dynamically stable provided $|\beta_2| < 1$. The immediate impact of a one unit change of x on y is β_3, and the long-run impact is $\beta_3/(1-\beta_2)$. Its straightforward (quasi maximum likelihood) estimator is $\hat{\beta}_3/(1-\hat{\beta}_2)$, where $\hat{\beta}_2$ and $\hat{\beta}_3$ can in fact be obtained by OLS if we condition on the observed value of y_0. Despite the restriction $|\beta_2| < 1$ on the DGP, the unrestricted OLS estimator $\hat{\beta}_2$ is such that $1 - \hat{\beta}_2$ has a nonzero density at argument zero, which may differ from zero substantially when β_2 is close to unity and σ not very small.

We shall illustrate the above in program mcs33.prg, in which we generate the series x_i as an arbitrary simple random walk process. Hence, variables y and x will both be nonstationary, or I(1), i.e., integrated of order one. Though, because u is I(0) the variables y and x are cointegrated, and $\beta_3/(1-\beta_2)$ is the coefficient of the cointegration vector. In the DGP of the program we will impose $\beta_3/(1-\beta_2)=1$. We simply take $\beta_1=0$ and $y_0=0$ and generate $u_i \sim NIID(0,\sigma^2)$. The program is:

```
'mcs33.prg:  MCS of an ARX(1) model
!R=10000          'number of Monte Carlo replications
!n=20             'sample size of the regressions
workfile f:\MCS\mcs33.wf1 u 0 !n
smpl 0 0
genr y=0
genr x=0
smpl 1 !n
rndseed 999
genr x=x(-1)+nrnd
!beta1=0
!beta2=0.5
!beta3=1-!beta2
!sigma=0.25
rndseed 9876543210
matrix (!R,5) simres
for !rep=1 to !R
  genr u = !sigma*nrnd
  genr y = !beta1 + !beta2*y(-1) + !beta3*x + u
  equation eq1.ls y c y(-1) x
  simres(!rep,1)=eq1.@coefs(2)
  simres(!rep,2)=eq1.@stderrs(2)
  simres(!rep,3)=eq1.@coefs(3)
  simres(!rep,4)=eq1.@stderrs(3)
  simres(!rep,5)=eq1.@R2
next
simres.write f:\MCS\mcs33sim.txt
```

```
workfile f:\MCS\mcs33sim.wf1 u 1 !R
read f:\MCS\mcs33sim.txt b2 seb2 b3 seb3 R2
genr tb2=(b2-!beta2)/seb2
genr rejtb2r=tb2>@qnorm(0.95)
genr rejtb2l=tb2<@qnorm(0.05)
genr TM=b3/(1-b2)
```

Note that variable x is kept fixed over all the replications. The consequences of that will be discussed later. We first examine the case $\beta_2 = 0.5$ and choose $\sigma = 0.25$ which yields an average coefficient of determination R^2 over the replications of 0.954. The histograms presented in Figure 3.5 show a very mild negative bias in $\hat{\beta}_2$ and virtually no bias in $\hat{\beta}_3$. The negative bias and skewness of $\hat{\beta}_2$ explain the displacement of the null-distribution of the test for β_2, which leads to estimated rejection probabilities at nominal significance level 5% of 8.0% against left-hand side alternatives and 4.1% against right-hand side alternatives. The bias and the distortions would be worse for larger σ and either better or worse for different x_i series, but milder for larger n. Remarkably, the estimator of the total multiplier (TM) seems to be centered around its true value 1 reasonably well. However, for larger σ and for larger R one would note most probably just a few extreme estimates. We shall next examine the behavior when $\beta_2 = 0.9$,

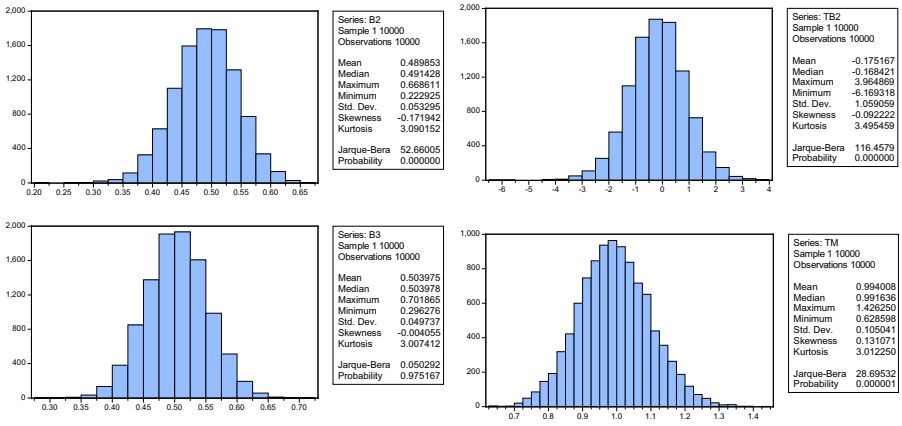

Fig. 3.5 ARX(1) results for $\beta_2 = 0.5$, $\sigma = 0.25$, and $n = 20$.

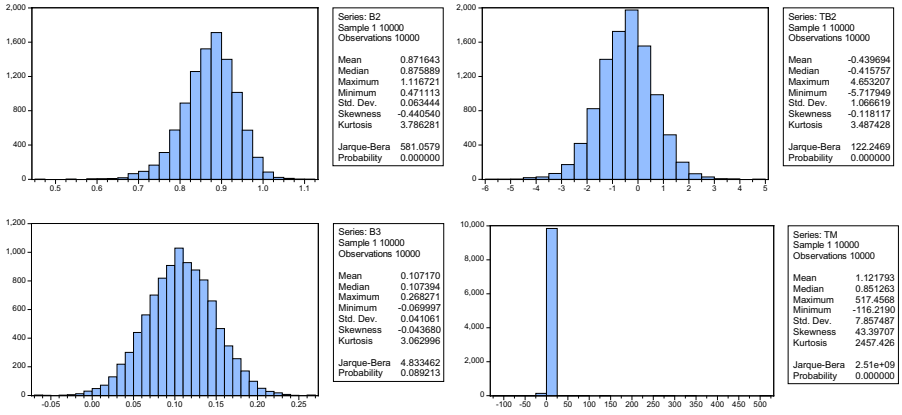

Fig. 3.6 ARX(1) results for $\beta_2 = 0.9$, $\sigma = 0.15$, and $n = 20$.

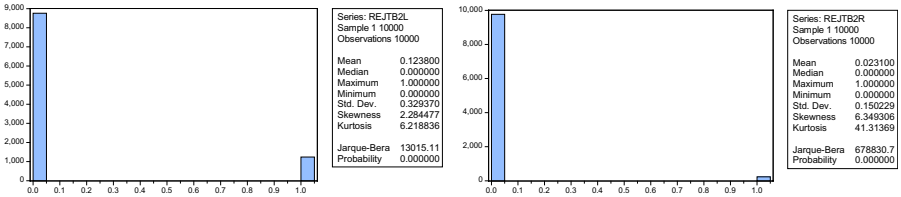

Fig. 3.7 Estimated actual significance levels for $H_0 : \beta_2 = 0.9$, $\sigma = 0.2$ at $n = 20$.

see Figure 3.6. We reduce the value of the disturbance standard deviation to $\sigma = 0.2$, which yields now an average R^2 of 0.951.

Figure 3.6 shows that the distributions of $\hat{\beta}_2$, $\hat{\beta}_3$ and the t-test deviate again from what first-order asymptotic theory predicts. Both $\hat{\beta}_2$ and $\hat{\beta}_3$ show a moderate but distinctive bias now. Figure 3.7 shows the very substantial size distortions of the standard inference methods when n is just 20, with actual significance levels of about 12% or about 2% instead of 5%. The histogram of TM in Figure 3.6 shows that its distribution is rather curious. In fact, of all the figures mentioned regarding the behavior of TM only median makes sense, because that estimates a quantile. All other figures on TM estimate moments which do not exist. Therefore these figures cannot be interpreted. The CLT does not apply, and therefore these functions of sample averages will not converge to fixed values. They remain random, irrespective of the magnitude of R. That positive outliers regarding TM were more prominent in our 10,000

draws is just a matter of chance; by choosing a different seed they might well turn out more negative. Hence, if we change R or change the random seed, we may obtain quite different outcomes. Note that from a probability theory point of view, the problem of the nonexistence of the moments of the estimator of TM, do occur irrespective of the actual values of β_2 and σ. However, its practical consequences only show up when β_2 is close to one and σ is not extremely small. When R is chosen small, the probability of the occurrence of any serious outliers might be so small, that the problem remains undetected.

The MCS estimate of the median of the distribution of the estimator of the total multiplier $\overrightarrow{\mathrm{MD}}(\widehat{\mathrm{TM}}) = 0.851$ suggests that in finite samples this median is not an unbiased estimator of the true value of TM. To be sure about that, we use (3.17) and evaluate $\overrightarrow{\mathrm{SD}}(\overrightarrow{\mathrm{MD}}(\widehat{TM})) = 0.005/\overrightarrow{f}_{\widehat{\mathrm{TM}}}(0.851)$. Figure 3.8 presents the EPDF of TM according to the standard approach of EViews. This suggests that the density in the median does not exceed 0.14, which would imply that the MCS estimate of the median has a standard error of about 0.03. However, due to the severe outliers this kernel density estimate appears to be very inaccurate. We modified variable TM [by the command: genr tmm=tm*(abs(tm)<5) + 5 * (tm >= 5) − 5 * (tm <= −5)] such that the tails of TMM are now less extreme, whereas this does not affect the mode and the median of their EPDF. From the density of TMM, see Figure 3.8, we find that $f_{\widehat{\mathrm{TM}}}(0.851) = 1.07$, so $\overrightarrow{\mathrm{SD}}(\overrightarrow{\mathrm{MD}}(\widehat{\mathrm{TM}})) = 0.0047$ and thus the true value of the median is in the range $[0.837, 0.865]$, which is far away from 1 indeed.

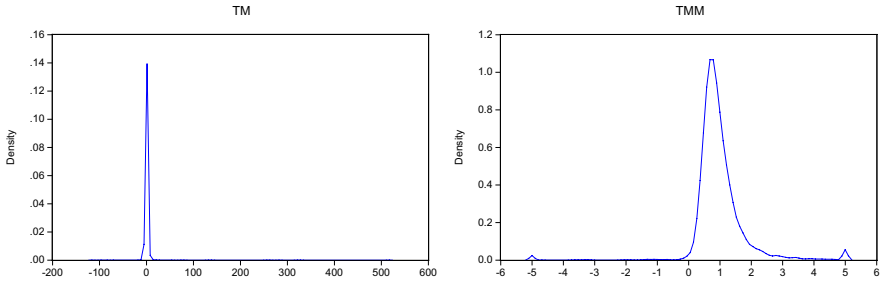

Fig. 3.8 Empirical PDF of TM and TMM: $\beta_2 = 0.9$, $\sigma = 0.1$ at $n = 20$.

Note that the results obtained by program mcs33.prg are conditional on the one and only realization of the x_t series that we generated. Is that what we want? Should we examine a few more typical random walk realizations for the x_t series, or should we generate a new x_t series in every replication? We will discuss such issues later in Section 5.

3.5 Results for Nonlinear Regression Models

The Box–Cox transformation is a one parameter nonlinear transformation of a positive variable which has the log and the linear transformation as two special cases. It is given by

$$B(x,\theta) = (x^\theta - 1)/\theta, \quad \theta \in \mathbb{R}, \;\; x \in \mathbb{R}^+. \tag{3.25}$$

For $\theta = 1$ this gives $x - 1$, and for $\theta = 0$ it is $\log x$ as follows from applying l'Hôpital's rule. However, evaluating it for $\theta = 0$ on a computer will give problems unless precautions are taken.[2] In the programs that follow we will use it first for a regressor variable, and estimate the resulting nonlinear regression model by Nonlinear Least Squares (NLS). Next we employ it to the dependent variable of a regression. Then estimation requires Maximum Likelihood (ML). Both NLS and ML are iterative estimation techniques in which, starting from an initial guess of the parameter values, a criterion function is optimized to find the estimates. When the DGP is a special case of the model specification these estimators have attractive asymptotic properties, but their behavior in finite samples is largely unknown. So, we will use MCS to shed some light on these properties and also on the effects of particular model specification errors.

3.5.1 An Analysis of Nonlinear Least-Squares

We examine a model that is nonlinear in its parameters $\beta_1, \beta_2, \theta_x$ and σ_u and has the form

$$y_i = \beta_1 + \beta_2 B(x_i,\theta_x) + u_i, \quad \text{with } u_i \sim IID(0,\sigma_u^2), \quad i = 1,\ldots,n. \tag{3.26}$$

[2] The present versions of the EViews programs in this section will derail when θ is chosen equal to zero, or when an (intermediate) estimate of θ will be zero (which has zero probability). This could be redressed by using the @recode function of EViews, see the next footnote.

NLS estimation involves the minimization over the parameter values of the sum of squared residuals

$$\sum_{i=1}^{n}[y_i - \beta_1 - \beta_2 B(x_i, \theta_x)]^2. \tag{3.27}$$

We focus on a case in which the series of explanatory variables is kept fixed in all the replications, although they were originally obtained by drawing them from a uniform distribution, as can be seen from program mcs34.prg.

```
'mcs34.prg: MCS with Box-Cox transformation of explanatory variable
!n=100
workfile f:\MCS\mcs34.wf1 u 1 !n
rndseed 12345
genr x=0.2+1.6*rnd   'artificial variable
genr logx=log(x)
!beta1=0
!beta2=1
!sigma=0.2
!thetax=0.3
genr xst=(x^!thetax-1)/!thetax
rndseed 9876543210
!R=10000
matrix (!R,8) simres
for !rep=1 to !R
   genr u=!sigma*nrnd     'iid Normal disturbances (0, sigma^2)
   genr y=!beta1 +!beta2*xst  + u
   param c(1) !beta1 c(2) !beta2 c(3) !thetax
   equation eq1.ls y-c(1)-c(2)*(x^c(3)-1)/c(3) 'NLS estimation
   simres(!rep,1)=eq1.@coefs(2)  'NLS estimate beta2
   simres(!rep,2)=(eq1.@coefs(2)-!beta2)/eq1.@stderrs(2)
   simres(!rep,3)=eq1.@coefs(3)  'NLS estimate thetax
   simres(!rep,4)=(eq1.@coefs(3)-!thetax)/eq1.@stderrs(3)
   equation eq2.ls y c xst 'unfeasible model estimated by OLS
   simres(!rep,5)=eq2.@coefs(2)  'OLS estimate beta2
   simres(!rep,6)=(eq2.@coefs(2)-!beta2)/eq2.@stderrs(2)
   equation eq3.ls y c x
   simres(!rep,7)=eq3.@coefs(2)
   equation eq4.ls y c logx
   simres(!rep,8)=eq4.@coefs(2)
next
simres.write f:\MCS\mcs34.txt
workfile f:\MCS\mcs34sim.wf1 u 1 !R
read f:\MCS\mcs34.txt b2 tb2 thetax tthetax b2e tb2e b2lin b2log
```

```
genr rejecttb2=abs(tb2)>@qnorm(0.975)
genr rejecttlxl=tthetax<@qnorm(0.05)
genr rejecttlxr=tthetax>@qnorm(0.95)
genr rejecttb2e=abs(tb2e)>@qtdist(0.975,!n-2)
```

To speed up the convergence we provide LS with the true parameter values as starting values. This is of course never possible in practice, but in MCS it helps to avoid convergence problems which one tackles in practice by trying various different starting values when needed.[3]

Figure 3.9 presents scatter plots of three OLS regressions from the final replication. The left-one is the unfeasible regression eq2 in which the true value of θ_x has been used. The other two are the misspecified regressions eq3 and eq4, respectively. Note that it is not easy to note from the latter two that the true dependence of y_i on the regressor is neither linear in x_i nor linear in $\log(x_i)$, especially not for the latter, because the chosen value of θ_x is closer to zero than to one. If we had chosen a smaller value of σ_u and/or realizations of the variable x_i that could be closer to zero as well as larger than 1.8 the true nonlinear nature of the relationship would have been more obvious from the scatters for eq3 and eq4.

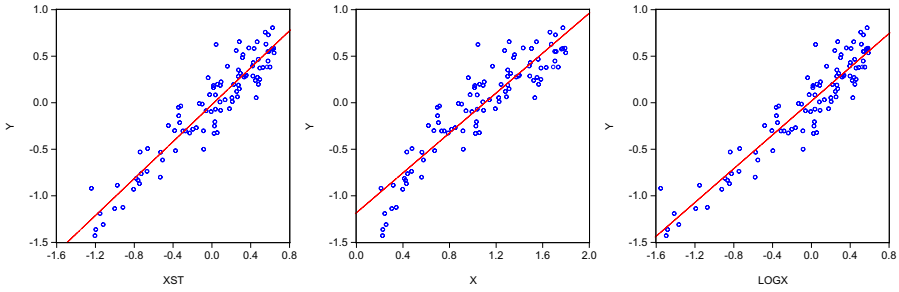

Fig. 3.9 Linear regression results obtained in the final replication of mcs34.prg.

[3] A more realistic program, using $\theta_x = .5$ as start-up value, would replace the "param line" and the next one by the three lines:

```
c(3)=0.5
equation eq0.ls y-c(1)-c(2)*(x^0.5-1)/0.5
equation eq1.ls y-c(1)-c(2)*@recode(x<> 0,(x^c(3)-1)/c(3),log(x))
```

This yields exactly the same results.

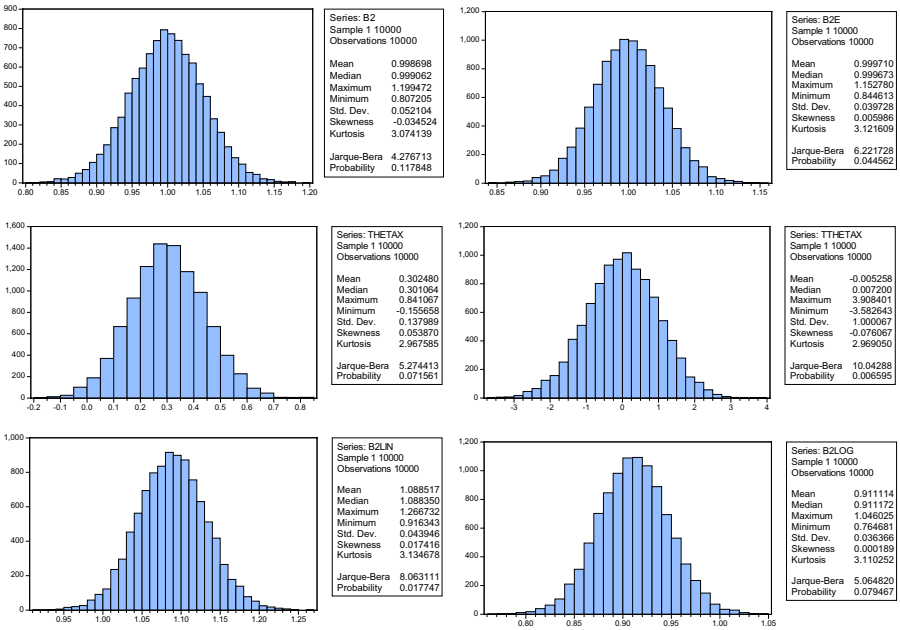

Fig. 3.10 Results from program mcs34.prg.

Figure 3.10 shows that the asymptotic normality and consistency of the NLS estimator for β_2 is obvious already at sample size $n = 100$. Comparing the values of $\overrightarrow{\mathrm{SD}}(\hat{\beta}_2)$ for B2 and B2E we also note that the efficiency loss with respect to the unfeasible estimator is only moderate. The NLS estimate of θ_x too does not show serious bias and has a distribution not seriously different from normal. Note that we have been "unlucky" with the Jarque–Bera test for the unfeasible OLS estimator, because we know that this is truly normal, given the normality of the disturbances. In NLS one obtains estimates of the variance of the coefficient estimates by numerically approximating the Hessian matrix of the residual sum of squares. Self-evidently, its quality will affect the distribution of test statistics. The MCS result for the histogram of the null distribution of the test for the value of θ_x seems already reasonably close to the standard normal, to which it converges asymptotically ($n \to \infty$). From the program we find rejection probability estimates at nominal significance level 5% of 0.0524 against the left-hand alternative and of

0.0470 against the right-hand alternative. For the two-sided NLS and the unfeasible OLS tests on the value of β_2 we find rejection frequency 0.0511 and 0.0500, respectively. Hence, our overall conclusion is that at $n = 100$ the asymptotic approximations work already extremely well for this type of nonlinearity. Note, though, that this may no longer be the case for a DGP with different x_i series, different parameter values and smaller sample size.

The histograms in Figure 3.10 on the β_2 estimates in the misspecified models show distributions which are not centered around the true value of $\beta_2 = 1$. This should not directly be classified as bias, because in a misspecified model the estimator converges to its so-called "pseudo true value." In this DGP we have that $\partial E(y_i \mid x_i)/\partial x_i^* = \beta_2$ is constant, but when we regress y_i on x_i OLS aims at estimating $\partial E(y_i \mid x_i)/\partial x_i = \beta_2 x_i^{\theta_x - 1}$, and when regressing on $\log(x_i)$ at $\partial E(y_i \mid x_i)/\partial \log(x_i) = \beta_2 x_i^{\theta_x}$. These vary with x_i. Because the x_i observations are evenly spread around 1 the result is that for $\theta_x = 0.3$ the slope with respect to x_i is larger than 1, but smaller with respect to $\log(x_i)$.

3.5.2 An Analysis of Maximum Likelihood

For a model in which the dependent variable is subjected to the Box–Cox transformation like

$$B(y_i, \theta_y) = \beta_1 + \beta_2 x_i + u_i, \quad \text{with } u_i \sim NIID(0, \sigma_u^2), \tag{3.28}$$

NLS estimation does not yield unique estimates. Adopting a particular form of the disturbance distribution enables to identify the parameters but requires employing ML. Expressing the PDF of the standard normal distribution for a variable ε_i as $\phi(\varepsilon_i) = (2\pi)^{-1/2} \exp(-\frac{1}{2}\varepsilon_i^2)$, we have for $u_i = \sigma_u \varepsilon_i$ that the Jacobian of the transformation is $1/\sigma_u$ yielding for u_i the density $f_{u_i}(u_i) = \sigma_u^{-1} \phi(u_i/\sigma_u)$. The transformation $u_i = B(y_i, \theta_y) - \beta_1 - \beta_2 x_i$ has Jacobian $\partial[B(y_i, \theta) - \beta_1 - \beta_2 x_i]/\partial y_i = y_i^{\theta - 1}$, so the density in terms of y_i is $f_{y_i}(y_i) = \sigma_u^{-1} \phi([B(y_i, \theta_y) - \beta_1 - \beta_2 x_i]/\sigma_u) y_i^{\theta_y - 1}$, which implies a contribution to the log-likelihood from observation i equal to

$$l_i = \log \phi((B(y_i, \theta_y) - \beta_1 - \beta_2 x_i)/\sigma_u) - \log \sigma_u + (\theta_y - 1) \log y_i. \tag{3.29}$$

ML involves the maximization of $\sum_{i=1}^{n} l_i$ over the parameter values. In program mcs35.prg the DGP has true value $\theta_y = 0.6$ and data are generated for the classic normal linear model $y_i^* = \beta_1 + \beta_2 x_i + u_i$, where $y_i^* = (y_i^{\theta_y} - 1)/\theta_y$, so $y_i = (\theta_y y_i^* + 1)^{1/\theta_y}$. Note that the chosen range of values for x_i and the values of the coefficients are slightly different in program mcs35.prg from those in program mcs34.prg.

```
'mcs35.prg: MCS with Box-Cox transformation of dependent variable
!n=100
workfile f:\MCS\mcs35.wf1 u 1 !n
rndseed 12345
genr x=0.1+1.8*rnd   'artificial variable
!beta1=-1
!beta2=1
!sigma=0.1
!thetay=0.6
rndseed 9876543210
!R=10000
matrix (!R,8) simres
for !rep=1 to !R
    genr u=!sigma*nrnd            'iid Normal disturbances (0, sigma^2)
    genr yst=!beta1 +!beta2*x + u
    genr y=(!thetay*yst+1)^(1/!thetay)
    genr logy=log(y)
    param c(1) !beta1 c(2) !beta2 c(3) !thetay c(4) !sigma 'ideal start-up values
    logl ml1    'these 6 lines perform Maximum Likelihood estimation ml1
    ml1.append @logl l 'variable l will contain the n contrib. to log-likelh.
    ml1.append ybc=(y^c(3)-1)/c(3)  'evaluation of Box-Cox transf. this iteration
    ml1.append u = ybc-c(1)-c(2)*x  'evaluation residuals in this iteration
    ml1.append l = log(@dnorm(u/c(4)))-log(c(4))+(c(3)-1)*log(y) 'log-llh. with Jcb.
    ml1.ml   'iterate until sum of log-likelihood contributions has been maximized
    simres(!rep,1)=ml1.c(2)    'ML estimate of beta2
    simres(!rep,2)=(ml1.c(2)-!beta2)/ml1.@stderrs(2) 't-test beta2
    simres(!rep,3)=ml1.c(3)    'ML estimate of thetay
    simres(!rep,4)=(ml1.c(3)-!thetay)/ml1.@stderrs(3)  't-test thetay
    equation eq1.ls yst c x  'infeasible model estimated by OLS
    simres(!rep,5)=eq1.@coefs(2)                      'OLS estimate beta2
    simres(!rep,6)=(eq1.@coefs(2)-!beta2)/eq1.@stderrs(2)      't-test
    equation eq2.ls y c x
    simres(!rep,7)=eq2.@coefs(2) 'OLS estimate beta2 in linear model
    equation eq3.ls logy c x
    simres(!rep,8)=eq3.@coefs(2) 'OLS estimate beta2 in log model
next
simres.write f:\MCS\mcs35.txt
workfile f:\MCS\mcs35sim.wf1 u 1 !R
read f:\MCS\mcs35.txt mlb2 tmlb2 thetay tthetay b2e tb2e b2lin b2log
genr rejecttmlb2=abs(tmlb2)>@qnorm(0.975)
genr rejectttyl=tthetay<@qnorm(0.05)
genr rejectttyr=tthetay>@qnorm(0.95)
genr rejecttb2e=abs(tb2e)>@qtdist(0.975,!n-2)
```

A practitioner would only observe y_i and x_i. If one knew the value of θ_y, one would calculate y_i^* and take that as the dependent variable and apply OLS. This technique, which is unfeasible in practice, is examined in eq1 in order to be able to see what the price is when not knowing θ_y. This price will be in terms of larger variance of coefficient estimates, and possible inaccuracies of asymptotic approximations in finite samples. We also examine the consequences of two types of misspecification. In eq2 we examine regressing y_i on an intercept and x_i and in eq3 using $\log(y_i)$ as the dependent variable. To speed up the convergence we again use the true parameter values as starting values in ML. Nevertheless, the iterations require substantial computing time and therefore running 10,000 replications takes a few minutes.

Figure 3.11 shows results from the final replication. Because θ_y is closer to 1 than to zero it is more likely that it is noticeable that the regression on x_i with $\log(y_i)$ as the dependent variable is misspecified than the one with y_i as the regressand.

From Figure 3.12 we note that the distribution of the ML estimator of β_2 is remarkably good, because its location and scale are almost similar to that of the unfeasible estimator which exploits the true value of θ_y. Also the ML estimator of θ_y shows hardly any bias. However, it is obvious that its standard error estimate must be very bad, because the null distribution of the coefficient test is both not well centered around zero and has a much larger standard deviation than one, which is its asymptotic value. Self-evidently, the asymptotic test must be heavily oversized in both tails. We found estimated type I error probabilities

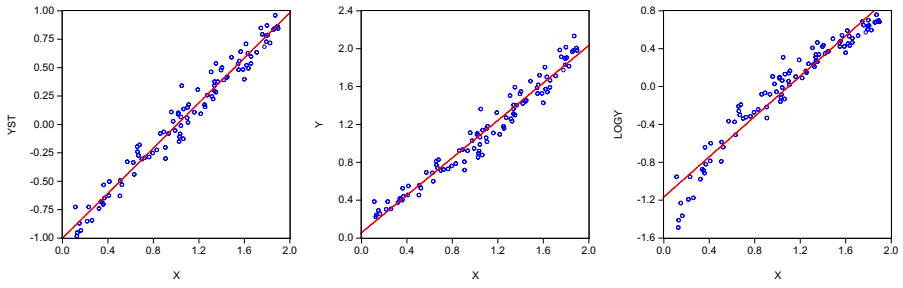

Fig. 3.11 Linear regression results obtained in the final replication of mcs35.prg.

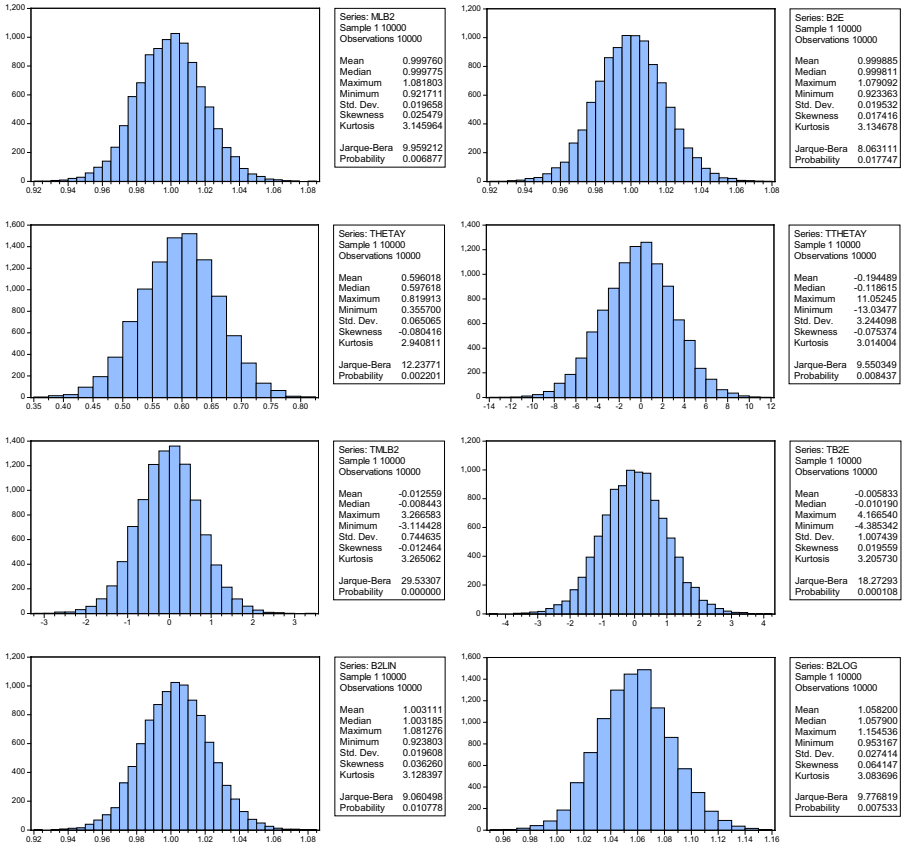

Fig. 3.12 Results from program mcs35.prg.

at the nominal 5% level of 0.3201 (left) and 0.2888 (right). While the standard error of the θ_y estimate is much too small on average, that of β_2 is much too large on average, because the null distribution of the coefficient test regarding β_2 has a standard error much smaller than 1. This yields a two sided type I error estimate of only 0.0114 at the 5% nominal level, whereas the exact but unfeasible test produces 0.0488.

From the various MCS results that were obtained in Sections 2 and 3 we have seen that in some cases asymptotic approximations are remarkably accurate, as in the classic linear regression model with non-normal errors. In many cases its coefficient tests seem to realize type I

error probabilities that are reasonably close to the aimed at nominal value. On the other hand, we found that consistent coefficient estimators in linear dynamic regression models may show substantial bias and skewness in finite samples. When their values are tested this can affect the actual type I error probabilities seriously, also under normality of the disturbances. For models that are nonlinear in their parameters a mixed picture emerges. In particular cases the null distributions of tests are extremely close to their asymptotic counterparts, whereas in other cases the actual type I error probabilities may be even five times as large, or five time as small, than desired. In the next section we will have a closer look into the origins of discrepancies between properties of tests in small and in large samples, and especially in what MCS can contribute to disclose and analyze such problems, and in developing possible remedies.

Exercises

1. Program mcs31.prg focusses on the accuracy and large-R asymptotic normality of MCS estimates of relatively small probabilities. What about estimating large probabilities, say 0.9, 0.99 or 0.999, by MCS?
2. Adapt program mcs31.prg such that you can establish MCS estimates of the probability that statistic $(\overrightarrow{p} - p)/[p(1-p)/R]^{1/2}$ is larger (smaller) than $+3(-3)$ for $R = 1{,}000, 5{,}000, 10{,}000$ and $p = 0.05, 0.025$. What rules of thumb do you distract regarding the required Monte Carlo sample size when estimating type I error probabilities of tests at nominal size 5% or 2.5%? Does this conform to the results from Table 3.1?
3. Indicate how you can calculate (without rerunning the program) the standard errors of the rejection probabilities presented in Table 3.3?
4. Change the value of !beta1 (the intercept) in program mcs32.prg and note that all results on OLS regarding β_2 and β_3 are unaffected. Next prove analytically that these results are invariant with respect to β_1 indeed.

5. Change the value of !beta2 and/or !sigma in mcs32.prg and examine the effects on the MCS estimates of the type I error probabilities.
6. Adapt program mcs32.prg such that you can verify the effect of disturbances with kurtosis much smaller than 3 by drawing the disturbances from a rescaled Uniform distribution. Use that when v is generated by rnd, then the components of $(v - 0.5) * \sqrt{12}$ are $IID(0,1)$.
7. Using the results given in Figure 3.3, provide confidence intervals for both the expectation and the median of $\hat{\beta}_2$ in the model with, and the one without the redundant regressor.
8. Consider ARX(1) model $y_i = \beta_1 + \beta_2 y_{i-1} + \beta_3 x_i + u_i$. Solve y_i in terms of x_i, $x_{i-1}, \ldots$ and establish the expression for the total multiplier.
9. Change program mcs11.prg or mcs12.prg such that the (pseudo) random variable is generated as one following a student distribution with 1 degree of freedom by changing the final line in genr t1=@rtdist(1). Explain why you don't find convergence of Mean for increasing values of n. Choose $n = 100$, 10,000, 1,000,000 and also change the rndseed value.
10. For the case of Figure 3.6, how many of the 10,000 TM estimates are in the interval [0,2]? By using the EViews smpl command examine the histogram of these realizations and note that due to scale problems the histogram of Figure 3.6 gives an odd impression of the actual smooth PDF of TM.
11. Rewrite program mcs33.prg such that the x_i series is a stationary AR(1) process and examine the effects on coefficient bias and test size distortions.
12. Use the delta-method to derive the asymptotic standard error of the TM estimator, and next examine any size distortions in testing $\beta_3/(1 - \beta_2) = 1$. Next compare with size distortions in testing simply the equivalent hypothesis $\beta_2 + \beta_3 = 1$.
13. How serious are the finite sample problems illustrated by program mcs33.prg when $n = 100$?
14. Examine by program mcs34.prg the dependence of the standard deviation of NLS estimators on n and on σ_u.

15. Examine by program mcs35.prg whether the asymptotic approximations work better when $n = 500$ (choosing a much lower value of R does already allow to come to useful conclusions).

4

Monte Carlo Analysis of Asymptotic Inference

Usually, the accuracy of inference in econometrics is hampered in finite samples due to the effects of so-called nuisance parameters. Although, provided the model is appropriately specified, estimators of parameters and estimators of their variance may be on target (consistent) when the sample is sufficiently large, in samples of the usual small or moderate size both may be biased. In many cases the magnitude of these biases can be shown to be $O(n^{-1})$ but their actual value is determined by a function of the true parameter values of the DGP and its deterministic components, usually too complex to be derived analytically. Of course this also affects test statistics, often being functions of estimators of parameters and their variance estimates. In the foregoing sections we have already seen how MCS can be used to assess bias and characteristics of the null distribution of test statistics. The latter enables to establish with great precision the actual type I error probability at the critical values suggested by asymptotic theory. In this section we will also consider the assessment of type II error probabilities by MCS and illustrate how MCS can be used to examine and possibly improve the general qualities of asymptotic inference procedures in finite samples.

For that purpose we will characterize test statistics and particular aspects of test procedures by introducing a helpful notation. Next we will indicate how appropriate critical values can be found when there are size problems for a particular test and DGP. However, instead of focussing on just one particular DGP with given parameter values, as we did in the previous sections, our major aim here is to indicate more generally what use can be made of MCS to examine and possibly improve the quality of inference when asymptotic tests are used for modelling and analyzing a particular type of relationship for which the actual parameter values are unknown.

We already saw that minor size problems occur when a classic regression model has nonnormal disturbances, although its coefficient estimates are still unbiased. Bias in the estimates of normal first-order autoregressive models with (or without) further exogenous regressors lead to more serious size problems in that class of models. Consider the ARX(1) model with k coefficients

$$y = (X \; y_{-1})\beta + u, \tag{4.1}$$

where $y = (y_1, \ldots, y_n)'$, $y_{-1} = (y_0, \ldots, y_{n-1})'$ and $(X \; y_{-1})$ is an $n \times k$ matrix. Let us assume for the moment that X is exogenous and that it is reasonable to condition on X and on the initial value y_0, and suppose that the disturbances are $u_i \sim IID(0, \sigma^2)$ and follow a yet unspecified distribution. Then the observed random vector y has some multivariate PDF that can be denoted as $f_y(y \mid X, y_0, \beta, \sigma)$. For testing $H_0 : C\beta = c$, where the $r \times k$ matrix C and the $r \times 1$ vector c contain known real constants, we use a scalar test statistic, say q_n, which should only depend on observables, i.e., $q_n = q(y, X, y_0, C, c)$. Let this have density $f_{q_n}(q; X, y_0, \beta, \sigma, C, c)$. When the null hypothesis is true density f_{q_n} specializes to $f^0_{q_n}(q; X, y_0, \eta, C, c)$. The latter represents the family of null distributions of the test, where η contains the nuisance parameters. Ideally the vector η is void, but in the present case it will contain elements from β and σ, which are generally unknown. This implies that the actual null distribution is unknown, even if we had specified the distribution of the disturbances u. So, fully appropriate critical values are not directly available. In what follows we use a notation similar to the one just introduced in order to represent a much

wider class of testing problems; testing general linear restrictions in the ARX(1) model just served as a relatively simple introduction.

4.1 General Characterization of Test Issues

Let, possibly after some (non)linear transformation of the parameter space,

$$q_n = q(y, X, y_0, \theta_{10}) \tag{4.2}$$

represent a test statistic for testing r simple restrictions on the parameters $\theta = (\theta_1', \theta_2')'$ of the model for an $n \times 1$ vector of dependent variables y. The restrictions tested are $H_0 : \theta_1 = \theta_{10}$, where vector θ_{10} contains r known real numbers; θ_2 represents the remaining parameters of the model for y, such as coefficients, parameters determining the distribution of the disturbances (including σ) and also, if required, those of the distribution of any endogenous or weakly exogenous explanatory variables of y. Formula (4.2) expresses that the test statistic can be calculated from the dependent variable y, the variables in X (which in case of endogenous explanatory variables may also include instrumental variables) the initial conditions y_0 (possibly a vector) and the values of the restrictions tested θ_{10}, which are all known. Its distribution will be determined by the full vector of parameters θ and therefore we denote its PDF as

$$f_{q_n}(q; X, y_0, \theta, \theta_{10}). \tag{4.3}$$

Under the null hypothesis this PDF specializes to

$$f^0_{q_n}(q; X, y_0, \eta, \theta_{10}), \tag{4.4}$$

where the vector of nuisance parameters η is θ_2 or a (possibly empty) subset of θ_2. In the classic normal linear regression model the null distribution is also free from initial conditions y_0, from the regressor matrix X and the actual value of θ_{10} (apart from their dimensions), and we simply have that the null distribution specializes to $f^0_{q_n}(q; r, n-k)$, which is the density of student's t (if $r = 1$) or otherwise Fisher's F. In what follows we focus both on the general case represented by (4.3) and (4.4), and on some particular special cases and their properties.

If we test H_0 against a (one-sided or two-sided) alternative H_1, whereas the statistic $q_n = q(y, X, y_0, \theta_{10})$ is constructed such that we reject H_0 in favor of H_1 for $q_n > Q_{1-\alpha}$, so $Q_{1-\alpha}$ is used as the critical value at nominal level α, then the test procedure has rejection probability

$$p(Q_{1-\alpha}; X, y_0, \theta, \theta_{10}) \equiv \Pr\{q_n > Q_{1-\alpha}\} = \int_{Q_{1-\alpha}}^{\infty} f_{q_n}(q; X, y_0, \theta, \theta_{10}) dq. \tag{4.5}$$

When H_0 is true this rejection probability is the actual type I error probability

$$\begin{aligned} p_0(Q_{1-\alpha}; X, y_0, \eta, \theta_{10}) &\equiv \Pr\{q_n > Q_{1-\alpha} \mid H_0\} \\ &= \int_{Q_{1-\alpha}}^{\infty} f^0_{q_n}(q; X, y_0, \eta, \theta_{10}) dq, \end{aligned} \tag{4.6}$$

which self-evidently depends in general on the value of the nuisance parameters η, and on X, y_0, and θ_{10}.

4.1.1 Pivots and (non)Similar Tests

When the distribution of a statistic does not depend on unknown parameters it is called a pivot. Obviously, it is possible and in fact very easy to simulate a pivot on a computer. When the null distribution of a test statistic does not depend on the unknown parameters η then it is pivotal, and we may denote its density as $f^{0*}_{q_n}(q; X, y_0, \theta_{10})$, where the symbol * highlights that no nuisance parameters are involved. When this pivotal density is well-behaved then for any $0 < p < 1$ the pth quantile $Q^*_p = Q^*_p(X, y_0, \theta_{10})$ is defined implicitly by

$$\int_{-\infty}^{Q^*_p} f^{0*}_{q_n}(q; X, y_0, \theta_{10}) dq = p, \tag{4.7}$$

and in principle this can be obtained easily, either by analytical or by experimental methods. If a pivotal null distribution depends just on the dimensions of X and θ_{10} only (hence, not on the actual elements of X and/or y_0), its quantiles can easily be tabulated; otherwise, they have to be obtained by integration (or by simulation, as we shall see) case by case. When a test has a pivotal null distribution and

$Q^*_{1-\alpha} = Q^*_{1-\alpha}(X, y_0, \theta_{10})$ is used as the critical value then its type I error probability is

$$p^*_0(Q^*_{1-\alpha}) \equiv \Pr\{q_n > Q^*_{1-\alpha} \mid H_0\} = \int_{Q^*_{1-\alpha}}^{\infty} f^{0*}_{q_n}(q; X, y_0, \theta_{10}) dq = \alpha. \tag{4.8}$$

Such a test procedure, which has constant rejection probability under the null, is called a similar test. When used with critical value $Q^*_{1-\alpha}$ it has actual type I error probability $p^*_0(Q^*_{1-\alpha})$ equal to the chosen nominal significance level α, irrespective of the values of n, X, y_0, θ_{10} and of the true value of θ in the DGP, which we will indicate by θ^0 below.

In most econometric models null distributions of tests are generally nonpivotal, giving nonsimilar tests, and type I error probabilities that cannot be controlled easily in finite samples, because they depend on nuisance parameters which have unknown values.

4.1.2 Asymptotic Tests

Let q_n be a test statistic that under the null hypothesis may be nonpivotal but has an asymptotic null distribution which is pivotal. This means that its asymptotic null distribution does not depend on unknown parameter values. This is the case for most of the test statistics commonly used in econometrics, and such tests are simply called asymptotic tests, which is shorthand for asymptotically similar and therefore asymptotically valid tests. Under the null hypothesis their asymptotic distribution is usually invariant with respect to X, y_0, and θ_{10} too. Hence $f^0_{q_n}(q; X, y_0, \eta, \theta_{10})$ converges for $n \to \infty$ to $f^{0*}(q; r)$, which has quantiles $Q^{*\infty}_p = Q^{*\infty}_p(r)$. Usually $f^{0*}(q; r)$ follows the $\chi^2(r)$ distribution (though when $r = 1$ it may be formulated such that it has a standard normal distribution). If one wants to test at nominal significance level α, then the critical values to be employed can simply be obtained from the appropriate table. Asymptotic p-values can be calculated by the computer by evaluating the CDF of the asymptotic null distribution for the observed empirical value $\hat{q}_n$ of the test statistic, giving

$$p^\infty(\hat{q}_n) \equiv \int_{\hat{q}_n}^{\infty} f^{0*}(q; r) dq. \tag{4.9}$$

Note, though, that this asymptotic p-value will generally differ from the actual p-value, which is

$$p_{q_n}(\hat{q}_n) \equiv \Pr\{q_n > \hat{q}_n \mid H_0\} = \int_{\hat{q}_n}^{\infty} f^0_{q_n}(q; X, y_0, \eta, \theta_{10}) dq. \tag{4.10}$$

As we have seen already, asymptotic tests may involve actual type I error probabilities which deviate from the nominal level, because

$$\begin{aligned} p_0(Q^{*\infty}_{1-\alpha}; X, y_0, \eta, \theta_{10}) &\equiv \Pr\{q_n > Q^{*\infty}_{1-\alpha} \mid H_0\} \\ &= \int_{Q^{*\infty}_{1-\alpha}}^{\infty} f^0_{q_n}(q; X, y_0, \eta, \theta_{10}) dq \end{aligned} \tag{4.11}$$

will generally differ from α, although for $n \to \infty$ it will converge to α. Hence, although in finite samples this test procedure is nonsimilar, asymptotically it is similar.

4.1.3 Exact Tests

When a critical value

$$\bar{Q}_{1-\alpha} = \bar{Q}_{1-\alpha}(X, y_0, \theta_{10}) \tag{4.12}$$

can be found, such that for any value of the nuisance parameters

$$\begin{aligned} p_0(\bar{Q}_{1-\alpha}; X, y_0, \eta, \theta_{10}) &\equiv \Pr\{q_n > \bar{Q}_{1-\alpha} \mid H_0\} \\ &= \int_{\bar{Q}_{1-\alpha}}^{\infty} f^0_{q_n}(q; X, y_0, \eta, \theta_{10}) dq \leq \alpha, \end{aligned} \tag{4.13}$$

then the test with rejection region $q_n > \bar{Q}_{1-\alpha}$ is called an exact test. Hence, an exact test rejects a true null hypothesis with a probability never exceeding the chosen nominal significance level. In that way it exercises limited control over the type I error probability. Self-evidently, a similar test directly enables exact inference, simply by exploiting $Q^*_{1-\alpha}$, and thus it has full control over its type I error probability. Note, though, that an under the null hypothesis nonpivotal test statistic which requires $\bar{Q}_{1-\alpha}$ to achieve exactness yields as a rule a nonsimilar test procedure, whereas on the other hand most nonsimilar tests are not exact when they use an asymptotic critical value $Q^{*\infty}_{1-\alpha}$ instead of $\bar{Q}_{1-\alpha}$.

In general, if a test is exact but nonsimilar, it will often be conservative, meaning that its actual rejection probability of a true null hypothesis is smaller than the nominal level. This will usually have a detrimental effect on the type II error probability of the test, as we will show below.

4.1.4 Modified Asymptotic Tests

Quite often the test statistics of asymptotic tests are actually not formulated and evaluated on the basis of their normal or $\chi^2(r)$ asymptotic null distributions. When $r = 1$ one often refers to the student distribution with a number of degrees of freedom equal to some positive integer number d, where $d = O(n)$; for instance, $d = n - k$, where k is the number of regressors. Asymptotically this does not matter, because for $n \to \infty$ this student distribution converges to the standard normal distribution. Nevertheless, in finite samples such modifications will usually lead to often minor (but occasionally, when n is very small, substantial) differences in rejection probabilities (as we will demonstrate below). Something similar may occur for tests where $r > 1$. Here often the χ^2 form of the test statistic is divided by r. Possibly it is also multiplied by d/n, which has an effect in finite samples, but not for $n \to \infty$ when $d = n - O(1)$. This modified statistic is then compared with critical values from the $F_{r,d}$ distribution. Asymptotically this does not matter either, because the $F_{r,d}$ distribution is the ratio of independent statistics $\chi^2(r)/r$ and $\chi^2(d)/d$. Because the denominator has expectation 1 and variance $2/d$, for $d \to \infty$ it has probability limit 1, so that $rF_{r,d}$ is asymptotically equivalent with the $\chi^2(r)$ distribution. However, for finite n these asymptotically equivalent test procedures will have different rejection probabilities. Such modifications (illustrated below) are attempts to mitigate size problems of asymptotic tests. How successful such an attempt is can be examined by MCS.

4.1.5 Three Types of Asymptotic Tests

Still considering statistic q_n to test $H_0 : \theta_1 = \theta_{10}$ in a model with parameters θ, we now explicitly take into account that the actual DGP is

supposed to have fixed true parameter values $\theta^0 = (\theta_1^{0\prime}, \theta_2^{0\prime})'$. By writing $q_n(\theta_{10}, \theta^0)$ for q_n we make explicit that in practice the test statistic will involve the tested value θ_{10} and that its distribution depends on the true values θ^0. The null distribution of the test static is pivotal if for $\theta_{10} = \theta_1^0$ the density of q_n is invariant with respect to θ_2^0 and thus has quantiles $Q_p^*(\theta_1^0)$. In practice, they will then often be invariant with respect to θ_1^0 (and X, y_0) too, but if not that would not necessarily be problematic, because under the null hypothesis these are known. Statistic $q_n(\theta_{10}, \theta^0)$ has a nonpivotal null distribution if it has quantiles $Q_p(\theta_1^0, \theta_2^0)$ which are not invariant with respect to θ_2^0. To be sure that the type I error probability will not exceed $100\alpha\%$, and still assuming that H_0 is rejected for large values of q_n, an observed value of q_n should then be compared with the critical value

$$\bar{Q}_{1-\alpha}(\theta_1^0) \equiv \max_{\theta_2} Q_{1-\alpha}(\theta_1^0, \theta_2), \tag{4.14}$$

because the actual value θ_2^0 is unknown. Assuming that $Q_{1-\alpha}(\theta_1^0, \theta_2)$ really varies with θ_2 this means that the test procedure that employs critical value $\bar{Q}_{1-\alpha}(\theta_1^0)$ will usually, if the maximum in (4.14) is not attained in θ_2^0, be conservative, implying that its actual type I error probability will be smaller than α. At any rate

$$\bar{Q}_{1-\alpha}(\theta_1^0) \geq Q_{1-\alpha}(\theta_1^0, \theta_2^0). \tag{4.15}$$

For the case of a nonsimilar test we shall now distinguish three different test procedures employing the same test statistic q_n for testing $\theta_1 = \theta_{10}$, but using three different critical values, namely $Q_{1-\alpha}(\theta_1^0, \theta_2^0)$, $\bar{Q}_{1-\alpha}(\theta_1^0)$ and $Q_{1-\alpha}^{\infty}(\theta_1^0)$. By the latter we denote the quantile of the asymptotic null distribution which is assumed to be pivotal. Note that the test using $Q_{1-\alpha}(\theta_1^0, \theta_2^0)$ is not operational, because in practice θ_2^0 is unknown. Also the test using $\bar{Q}_{1-\alpha}(\theta_1^0)$ is not easy to apply in practice, because evaluating (4.14) analytically may prove to be too cumbersome. However, later we will show that one can get close to this procedure by using advanced Monte Carlo methods. The third procedure is what one usually does in practice, possibly using a modification of the asymptotic test with good reputation. Nevertheless, using $Q_{1-\alpha}^{\infty}(\theta_1^0)$ will always involve some loss of control over actual type I error probability, which may also undermine the power of the test, as we shall see.

The type II error probability (not rejecting a false null) of the three test procedures does of course depend on the critical value that they use. For the first procedure it is $F_{q_n}(Q_{1-\alpha}(\theta_1^0, \theta_2^0))$, where by $F_{q_n}(\cdot)$ we indicate the CDF of q_n which is in general determined by X, y_0, θ^0, and θ_{10}. For the second procedure it will be $F_{q_n}(\bar{Q}_{1-\alpha}(\theta_1^0))$. From (4.15) we obtain the inequality

$$F_{q_n}(\bar{Q}_{1-\alpha}(\theta_1^0)) \geq F_{q_n}(Q_{1-\alpha}(\theta_1^0, \theta_2^0)). \tag{4.16}$$

For the asymptotic test it is $F_{q_n}(Q^{\infty}_{1-\alpha}(\theta_1^0))$, which could be either larger or smaller than the type II error probability of the other two procedures. However, when it is smaller this cannot be interpreted as having larger power, because its type I error probability may be larger too.

4.1.6 Some Further Test Issues

To correct in a MCS study for the effects of different type I error probabilities when comparing the type II error probabilities of alternative asymptotic test procedures, one often calculates the so-called "size corrected power" or its complement "size corrected type II error probabilities." These are obtained by using as critical value MCS estimates of either $\bar{Q}_{1-\alpha}(\theta_1^0)$ or $Q_{1-\alpha}(\theta_1^0, \theta_2^0)$. It may be very computationally involved to obtain the former quantile. However, its attraction is that it respects that in practice one does not know the true value of the nuisance parameters. In the latter, which is more popular, one exploits that in the MCS experiments the value of θ_2^0 is actually known. Note, however, that a test statistic which has the best power after such size correction is not necessarily the one to be recommended for use in practice, because if one uses that statistic in practice with its asymptotic critical value $Q^{\infty}_{1-\alpha}$, it is again affected by lack of size control, which jeopardizes the accuracy of inference. In extreme cases, when the actual type I error probability is very high, a large test statistic could both be due to either a true or to a false null hypothesis. And when the actual type I error probability is much smaller than the nominal significance level then a very small value of the test statistic could also be due both to a true and to a false null. Hence, only test procedures should be taken into consideration about which one knows that their actual type

I error probability, when used in practice, may deviate only mildly from the nominal significance level.

Various of the issues addressed above will be illustrated now. We mainly focus on estimating rejection probabilities of test statistics either under the null hypothesis or under a few specific forms of alternative hypotheses. The results should be useful not just to understand the issues addressed above, but also with respect to the following. In the above, we assumed that the actual DGP is a special case of the model considered. In practice, however, test procedures will be used very often in situations where neither the null nor the alternative hypothesis is strictly true. For instance, the DGP may involve second order serial correlation, whereas a test is applied for first-order serial correlation. Or more seriously and relevant for practitioners, one may apply a test for serial correlation in a model that lacks some regressors that have nonzero coefficients in the DGP. These omitted regressors have either been simply overlooked or were not available, or one did not realize that the functional form requires powers or other transformations of some of the included regressors, or it lacks interaction terms or particular lagged variables. Practitioners should always be aware when they interpret the outcomes of test statistics that a value of the test statistic with a high (or at least not a small) p-value although corroborating the null hypothesis, certainly does not imply validity. It could as well be the case that the test is either undersized (rejects too infrequently anyhow) and/or suffers from lack of power, either because the sample is too small, or the incorrectness of the null is rather modest, or it may test against the wrong alternative, because the actual DGP is not covered by the alternative hypothesis of the test. In addition, the rejection of a null hypothesis does not imply the truth of the alternative against which it was tested. These phenomena can easily be illustrated by MCS experiments.

4.2 Size Correction and Power Comparisons

In the section above we already made various remarks on size correction and the role that it plays to enable a fair comparison of the powers of alternative test procedures. Here we will provide a few illustrative

examples of some of the issues involved. However, we will not come up with a fully fledged example of size correction and power comparison because of reasons already indicated above. A test procedure can only safely be used in practice if it allows reasonable control over its type I error probability. Hence, only when techniques for size correction are of a nature such that they can be employed successfully in practice, then they are really useful. Such techniques (Monte Carlo testing and also bootstrapping) will be discussed in Section 6. Occasionally, however, crude asymptotic tests do already achieve reasonable control over type I errors, possibly after some type of finite sample modification.

In this section we will illustrate what kind of size correction can be established by critical values of the type $\overrightarrow{Q}_{1-\alpha}(\theta_1^0, \theta_2^0)$ and the limited usefulness of MCS based power comparisons that they allow. At the same time we will be able to illustrate that in a few particular situations the size problems seem generally mild, whereas they can be quite serious in other cases. However, we repeat: Test procedures with serious size problems which can only be overcome in a simulation study (where nuisance parameter values are known) but not in practice, should be disqualified for use in practice, and so it is not very useful to let them participate in power comparison studies. So, because such studies should only include procedures with reasonable control over type I errors, then there is in fact no serious need for size correction in MCS either.

4.2.1 An Illustration of Size Correction

We return to program mcs32.prg in which we examined the effect of skew disturbances in a classic linear regression. We shall analyze what appropriate 5% one-sided critical values would be for the tests on β_2 and β_3 when $n = 20$ as obtained from the simulation with only $R = 2{,}500$ replications. For $p = 0.05, 0.95$ we find $\lceil Rp \rceil = 125$ and 2,375, respectively. Sorting the q_n series, we find from q_n^* for TB2[1] that $\overrightarrow{Q}_{0.05} = -1.93$ and $\overrightarrow{Q}_{0.95} = 1.51$, and for TB3 that $\overrightarrow{Q}_{0.05} = -1.54$ and $\overrightarrow{Q}_{0.95} = 1.90$. These deviate substantially from the corresponding

[1] In the workfile window give Proc, Sort Current Page ... and provide as sort key tb2 and make sure that you sort in ascending order.

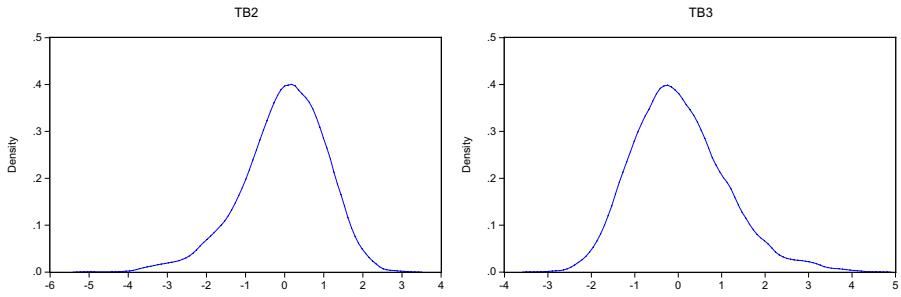

Fig. 4.1 EPDF of null distributions of TB2 and TB3 of mcs32.prg.

quantiles from the normal and the student with 17 degrees of freedom, which are ± 1.645 and ± 1.74, respectively.

To assess the standard errors of these quantile estimates we have to establish the relevant $\overrightarrow{f}_q(\overrightarrow{Q}_p)$, which should better be taken too large than too small in order to prevent being too optimistic about the accuracy of $\overrightarrow{Q}_p$. Both from the histograms in Figure 3.3 and from the kernel density estimates in Figure 4.1 we can see that in all four cases $\overrightarrow{f}_q(\overrightarrow{Q}_p)$ is about 0.1. Because $[p(1-p)/R]^{1/2} = 0.00436$ (for $p = 0.05, 0.95$) for all four examined cases we find that $\overrightarrow{SD}(\overrightarrow{Q}_p)$ is about 0.044. Hence, the half-width of a confidence interval will be about 0.13 or almost 10%. From this we learn that a MCS sample size of 2,500 seems rather small to establish size corrected critical values with reasonably high precision. However, in Section 6 we will return to this issue when discussing pivotal Mone Carlo tests.

4.3 Analysis of Various Misspecification Tests

In this section, we collect a few more illustrations of asymptotic test procedures which can be modified in a straight-forward way. By MCS we study and compare for very specific DGP's their control over type I error probabilities and produce some results on test power.

4.3.1 The RESET Test in Classic Normal Regression

The RESET test is a rather straightforward device to check the adequacy of the functional form of a linear regression specification with IID

disturbances, without being very explicit about the alternative against which one tests. The procedure, which can be rationalized by casting it into the concept of a Lagrange multiplier test on the linearity of a nonlinear regression specification, involves first estimating the regression model under the null hypothesis, which is

$$y_i = x_i'\beta + u_i, \tag{4.17}$$

where x_i contains k regressors and $u_i \sim IID(0,\sigma^2)$ for $i = 1,\ldots,n$. One applies OLS and obtains residuals $\hat{u}_i$ and fitted values $\hat{y}_i = y_i - \hat{u}_i$. Next, the test involves testing the significance in the usual way of the $m \geq 1$ extra regressors in the auxiliary model

$$y_i = x_i'\beta + \gamma_1\hat{y}_i^2 + \cdots + \gamma_m\hat{y}_i^{m+1} + u_i \tag{4.18}$$

by an $F_{m,n-k-m}$ test, or if $m = 1$ by the habitual t_{n-k-1} test.

It can be derived that the test is asymptotically valid, so for large samples the critical values of the χ^2 distribution, or after a modification by the F, allow size control, provided that the regressors x_i are predetermined, which requires $u_i \mid x_i \sim IID(0,\sigma_u^2)$. In finite samples the null distribution may deviate from the asymptotic null distribution, especially when the disturbances are highly nonnormal, but also because the contemporaneous correlation between u_i and $\hat{y}_i^j$ for $j = 2,\ldots,m$ may cause finite sample problems.

In program mcs41.prg we consider the tests for $m = 1$ and for $m = 2$, both their crude asymptotic version and a modification, in a model where $k = 2$ (intercept and a smoothly cycling regressor x_i). First the data are generated according to the model of the null hypothesis with disturbances drawn from the Uniform distribution. Next a regression is run in which the regressor is $\log(x_i)$ and we examine how frequently the RESET test discovers that the functional form of this model is inappropriate.

```
'mcs41.prg: MCS of RESET test
!n=40
workfile f:\MCS\mcs41.wf1 u 1 !n
genr i=@trend(0)
genr ii=i-20*@floor((i-1)/20)
!averagex=1
genr x=!averagex+0.5*sin(2*@acos(-1)*(ii-1)/20)          'exogenous x with cycle
```

```
genr logx=log(x)
!beta1=0
!beta2=1
!sigma=0.2
!factor=!sigma*@sqrt(12)
!R=10000
matrix (!R,6) simres
rndseed 9876543210
for !rep=1 to !R
   genr u=!factor*(rnd-0.5)
   genr y=!beta1+!beta2*x+u
   equation eq1.ls y c x
   genr yhat=y-resid
   equation eq2.ls y c x yhat^2
   simres(!rep,1)=@sqrt(!n/(!n-3))*eq2.@coefs(3)/eq2.@stderrs(3)
   simres(!rep,2)=eq2.@coefs(3)/eq2.@stderrs(3)
   equation eq3.ls y c x yhat^2 yhat^3
   simres(!rep,3)=!n*(eq1.@ssr-eq3.@ssr)/eq3.@ssr
   simres(!rep,4)=((!n-4)/(2*!n))*simres(!rep,3)
   equation eq4.ls y c logx
   genr yhat=y-resid
   equation eq5.ls y c logx yhat^2
   simres(!rep,5)=eq5.@coefs(3)/eq5.@stderrs(3)
   equation eq6.ls y c logx yhat^2  yhat^3
   simres(!rep,6)=((!n-4)/2)*(eq4.@ssr-eq6.@ssr)/eq6.@ssr
next
simres.write f:\MCS\mcs41sim.txt
workfile f:\MCS\mcs41sim.wf1 u 1 !R
read f:\MCS\mcs41sim.txt r10 r10m r20 r20m r11m r21m
genr rejr10=abs(r10)>@qnorm(0.975)
genr rejr10m=abs(r10m)>@qtdist(0.975,!n-3)
genr rejr20=r20>@qchisq(0.95,2)
genr rejr20m=r20m>@qfdist(0.95,2,!n-4)
genr rejr11m=abs(r11m)>@qtdist(0.975,!n-3)
genr rejr21m=r21m>@qfdist(0.95,2,!n-4)
```

In Figure 4.2 graphs are presented for $n = 40$ for explanatory variables x and $\log x$, for the drawings of the disturbance u in the final replication and the resulting dependent variable y. Note that for the chosen range of values for the regressor x there is substantial multicollinearity with $\log(x)$.

To what degree the null distributions of the test statistics for $m = 1$ deviate from the normal and t_{37}, and for $m = 2$ from χ^2_2 and $F_{2,36}$, can be examined by investigating the histograms of the variables r10, r10m, r20, and r20m, but as far as size control is concerned just the tail areas

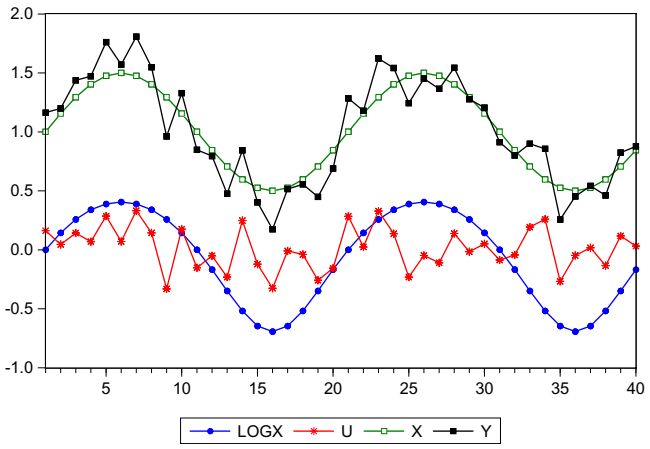

Fig. 4.2 Variables in final replication, $n = 40$.

Table 4.1. Rejection probabilities RESET for $\alpha = 0.05, n = 40$.

	REJR10	REJR10m	REJR11m	REJR20	REJR20m	REJR21m
$\overrightarrow{p}$	0.0659	0.0489	0.2986	0.0776	0.0492	0.2236
$\overrightarrow{\mathrm{SD}}(\overrightarrow{p})$	0.0025	0.0022	0.0046	0.0027	0.0022	0.0042

are relevant, summarized by the rejection frequencies $\overrightarrow{p}$ given below. For $n = 40$ these are given in Table 4.1.

We note that from $R = 10{,}000$ we cannot establish the actual significance level very precisely yet, but we do find that for this particular DGP and nonnormal disturbances there is certainly no serious difference between the nominal and the actual significance level provided one uses the modified test procedures. Using the crude asymptotic test leads to overrejection. Moreover, we find that the rejection probabilities when the null is false, although much larger than 5%, are disappointingly low, which must be due to the similarity in the patterns of x and $\log(x)$ as shown in Figure 4.2. Table 4.2 shows the results for $n = 200$.

From this we learn that even for a larger sample it makes sense to use a small sample modification to an asymptotic test and, although we are not in the classic normal linear regression model, to mimic standard t and F procedures as closely as possible. Using the modification, we

Table 4.2. Rejection probabilities RESET for $\alpha = .05, n = 200$.

	REJR10	REJR10m	REJR11m	REJR20	REJR20m	REJR21m
$\overrightarrow{p}$	0.0547	0.0520	0.9154	0.0603	0.0552	0.8580
$\overrightarrow{\mathrm{SD}}(\overrightarrow{p})$	0.0023	0.0022	0.0028	0.0024	0.0023	0.0035

notice that in the auxiliary linear regression model (4.18), which is nonnormal and nonclassic, the RESET test is not suffering from serious size distortions. Even for a sample size as large as 200 the probability of type II errors is still in the range of 8%–14% for this particular setting. Note that some estimates of the actual significance level at the larger sample size deviate more from the nominal value of 5% than those for the smaller sample size. Given the fact that in our DGP the regressor is "constant in repeated samples of size 20" we do suppose that the actual significance level will converge for n large monotonically to 5%. In that case the reason that our MCS results do not follow this pattern must be entirely due to the random nature of $\overrightarrow{p}$, which is in agreement with the values of $\overrightarrow{\mathrm{SD}}(\overrightarrow{p})$ indeed.

4.3.2 Analysis of Some Tests for Serial Correlation

Next we examine tests for serial correlation in a linear regression of the form (4.17) but we shall use a DGP of the ARX(1) type, hence regressor y_{i-1} is not exogenous. When the disturbances are IID y_{t-1} is predetermined, but when the disturbances show first-order serial correlation then y_{i-1} and u_i are jointly dependent and OLS estimators will be inconsistent. The presence of regressor y_{i-1} renders the Durbin–Watson test unsuitable, because the tabulated bounds on its critical values are only justified for models with exogenous regressors. We will examine whether we can establish that the Durbin–Watson test is improper in this model. We will also examine a first-order form of the Box–Pierce test, which is improper too in this model, and make comparisons with Lagrange multiplier versions of serial correlation tests, known as the Breusch–Godfrey implementation, which are asymptotically valid for ARX models. For the latter tests we again compare crude versions (with critical values from the normal and the χ^2 distribution) with modified versions (in which critical values for the t and F are

being used). The program is as follows:

```
'mcs42.prg: MCS on size of serial correlation tests in an ARX(1) model
!n=30
!warmup=100
!warmupn=!warmup+!n
workfile f:\MCS\mcs42.wf1 u 1 !warmupn
!beta1=0
!beta2=0.2
!beta3=1-!beta2
!sigma=0.05
!gamma=0.8
smpl 1 1
genr y=0
genr x=0
rndseed 12345
smpl 2 !warmupn
genr x=!gamma*x(-1)+@sqrt(1-!gamma^2)*nrnd
genr u=!sigma*nrnd
genr y=!beta1+!beta2*x+!beta3*y(-1)+u
smpl !warmup+1 !warmupn
!R=10000
matrix (!R,8) simres
rndseed 9876544210
for !rep=1 to !R
   genr u=!sigma*nrnd
   genr y=!beta1+!beta2*x+!beta3*y(-1)+u
   equation eq1.ls y c x y(-1)
   simres(!rep,1)=eq1.@dw
   genr resid1=resid
   genr resid1lag=resid1(-1)
   smpl !warmup+1 !warmup+1
   genr resid1lag=0
   smpl !warmup+1 !warmupn
   equation eq2.ls resid1 resid1lag
   simres(!rep,2)=(eq2.@tstats(1))^2
   equation eq3.ls resid1 c x y(-1) resid1lag
   simres(!rep,3)=eq3.@tstats(4)
   genr resid1lag2=resid1lag(-1)
   smpl !warmup+1 !warmup+1
   genr resid1lag2=0
   smpl !warmup+1 !warmupn
   equation eq4.ls resid1 c x y(-1) resid1lag resid1lag2
   simres(!rep,4)=!n*eq4.@r2
   simres(!rep,5)=((!n-5)/2)*(eq1.@ssr-eq4.@ssr)/eq4.@ssr
   genr resid1lag3=resid1lag2(-1)
   smpl !warmup+1 !warmup+1
```

```
   genr resid1lag3=0
   smpl !warmup+1 !warmupn
   equation eq5.ls resid1 c x y(-1) resid1lag resid1lag2 resid1lag3
   simres(!rep,6)=!n*eq5.@r2
   simres(!rep,7)=((!n-6)/3)*(eq1.@ssr-eq5.@ssr)/eq5.@ssr
   simres(!rep,8)=eq1.@r2
next
simres.write f:\MCS\mcs42sim.txt
workfile f:\MCS\mcs42sim.wf1 u 1 !R
read f:\MCS\mcs42sim.txt dw BP1 LMt1 LM2 LMF2 LM3 LMF3 R2
genr rejBP1=BP1>@qchisq(0.95,1)
genr rejLMt1L=LMt1<@qtdist(0.05,!n-4)
genr rejLMt1R=LMt1>@qtdist(0.95,!n-4)
genr rejLMt1=abs(LMt1)>@qnorm(0.975)
genr rejLM2=LM2>@qchisq(0.95,2)
genr rejLMF2=LMF2>@qfdist(0.95,2,!n-5)
genr rejLM3=LM3>@qchisq(0.95,3)
genr rejLMF3=LMF3>@qfdist(0.95,3,!n-6)
```

In program mcs42.prg the ARX(1) model has three explanatory variables: an intercept, one exogenous regressor, and the lagged dependent variable. Because of the latter there is a start-up problem which we have solved as follows. This procedure, which is often followed when generating autoregressive models, is known as "warming-up the DGP." We start with values of the series $x_1 = 0$ and $y_1 = 0$ equal to their unconditional mean. From there on we start to generate $x_t = \gamma x_{t-1} + v_t$ and $y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t$ for $t = 2, 3, \ldots$ where $v_t \sim NIID(0, 1 - \gamma^2)$ and $u_t \sim NIID(0, \sigma^2)$. Note that x_t will be a stationary AR(1) process with $\mathrm{Var}\, g(x_t) = 1$ only for large t, due to the nonrandom zero start-up. Also y_t will only get gradually on its stationary track. In the program we take the 100th drawings as the last presample value for the actual sample of $n = 30$ observations. We may assume that from observation 100 on, y_t has a variance very close to its stationary value. Moreover, now y_{t-1} and x_t will have the appropriate correlation, which is certainly not yet the case shortly after the initialization of the process.[2]

[2] Generating the AR(1) process for x_t such that it is stationary from the first observation on is actually quite simple, as was already indicated in program mcs23.prg. For y_t it is more cumbersome, but possible too, see Kiviet (1985).

The chosen value of the autoregressive coefficient of x_t makes it a rather smooth series. The regression coefficients in combination with the value of σ yield an average R^2 for the correctly specified model of 0.98 on average. We will not focus on the distribution of the OLS coefficient estimators and related tests, but on serial correlation tests.

For the Durbin–Watson bounds test at $n = 30$ and $k = 3$ the critical values against positive serial correlation at significance level 5% are $d^L = 1.28$ and $d^U = 1.57$. We can establish from the 10,000 drawings dw that for NIID disturbances $\overrightarrow{\Pr}(dw < 1.28) = 0.0090$ (0.0009) and $\overrightarrow{\Pr}(dw < 1.57) = 0.0655$ (0.0025), from which we find that the exact 5% critical value should be somewhere between d^L and d^U indeed, so the inappropriateness (asymptotic invalidity) of the test does not emerge from these results. However, below we will examine whether this perhaps occurs when increasing n.

The Box–Pierce test is also asymptotically invalid because of the occurrence of y_{i-1} as regressor. From Table 4.3, we see that in the present DGP its type I error probability does deviate from 5%, though not very seriously. All the other tests examined are asymptotically valid. We note that testing against second or third order serial correlation when using the crude (χ^2 type) version leads to a slightly more serious larger type I error probability than is the case for the F-type implementation. Using a two-sided t-type test for first order serial correlation leads to a type I error probability that is significantly above 5%, though by a margin that would not worry many practitioners. However, this deviation shows that the null distribution is affected by the nuisance parameters and gives rise to concerns that for different parameter values of the DGP the size problems may be much worse. Also, we find that this two-sided actual significance level is the result of one-sided significance levels which are seriously different from the nominal value. We find serious overrejection against negative serial correlation and underrejection against positive serial correlation.

When examining $n = 200$ in Table 4.4 we have $d^L = 1.75$ and $d^U = 1.79$ and find $\overrightarrow{\Pr}(dw < 1.75) = 0.0254$ (0.0016) and $\overrightarrow{\Pr}(dw < 1.79) = 0.0528$ (0.0022), so again we do not find that the asymptotic invalidity of the Durbin–Watson test in the ARX(1) model disqualifies it in finite samples. Either we need an even higher value of n to establish its failure,

Table 4.3. Rejection probabilities of serial correlation tests, $n = 30$.

	BP1	LM2	LM3	LMF2	LMF3	LMT1	LMT1L0	LMT1R
$\overrightarrow{p}$	0.075	0.073	0.063	0.059	0.052	0.067	0.096	0.014
$\overrightarrow{\mathrm{SD}}(\overrightarrow{p})$	0.003	0.003	0.002	0.002	0.002	0.003	0.003	0.001

Table 4.4. Rejection probabilities of serial correlation tests, $n = 200$.

	BP1	LM2	LM3	LMF2	LMF3	LMT1	LMT1L0	LMT1R
$\overrightarrow{p}$	0.047	0.049	0.049	0.048	0.048	0.046	0.065	0.031
$\overrightarrow{\mathrm{SD}}(\overrightarrow{p})$	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.002

or the actual size happens to be close to 5% for the present coefficient values. For the other tests we find that, also for the invalid BP test, all size problems are now absent or much more moderate than for the smaller sample size.

Next we adapt the program and obtain mcs43.prg in which the disturbances now show first-order serial correlation with serial correlation coefficient $\rho = 0.4$. The program is a straight-forward adaptation of mcs42.prg. However, now we have taken into account that in the simple AR(1) processes for x_t and u_t the warming-up problem can easily be solved by drawing the initial value immediately from the unconditional distribution instead of choosing a value equal to the unconditional expectation.

```
'mcs43.prg: MCS on power of serial correlation tests in an ARX(1) model
!n=30
!warmup=100
!warmupn=!warmup+!n
workfile f:\MCS\mcs43.wf1 u 1 !warmupn
!beta1=0
!beta2=0.2
!beta3=1-!beta2
!sigma=0.05
!rho=0.4
!gamma=0.8
rndseed 12345
smpl 1 1
genr y=0
genr x=nrnd
genr u=!sigma*nrnd
```

```
smpl 2 !warmupn
genr x=!gamma*x(-1)+@sqrt(1-!gamma^2)*nrnd
genr u=!rho*u(-1)+!sigma*@sqrt(1-!rho^2)*nrnd
genr y=!beta1+!beta2*x+!beta3*y(-1)+u
smpl !warmup+1 !warmupn
!R=10000
matrix (!R,8) simres
rndseed 9876544310
for !rep=1 to !R
   genr u=!rho*u(-1)+!sigma*@sqrt(1-!rho^2)*nrnd
   genr y=!beta1+!beta2*x+!beta3*y(-1)+u
   equation eq1.ls y c x y(-1)
   simres(!rep,1)=eq1.@dw
   genr resid1=resid
   genr resid1lag=resid1(-1)
   smpl !warmup+1 !warmup+1
   genr resid1lag=0
   smpl !warmup+1 !warmupn
   equation eq2.ls resid1 resid1lag
   simres(!rep,2)=(eq2.@tstats(1))^2
   equation eq3.ls resid1 c x y(-1) resid1lag
   simres(!rep,3)=eq3.@tstats(4)
   genr resid1lag2=resid1lag(-1)
   smpl !warmup+1 !warmup+1
   genr resid1lag2=0
   smpl !warmup+1 !warmupn
   equation eq4.ls resid1 c x y(-1) resid1lag resid1lag2
   simres(!rep,4)=!n*eq4.@r2
   simres(!rep,5)=((!n-5)/2)*(eq1.@ssr-eq4.@ssr)/eq4.@ssr
   genr resid1lag3=resid1lag2(-1)
   smpl !warmup+1 !warmup+1
   genr resid1lag3=0
   smpl !warmup+1 !warmupn
   equation eq5.ls resid1 c x y(-1) resid1lag resid1lag2 resid1lag3
   simres(!rep,6)=!n*eq5.@r2
   simres(!rep,7)=((!n-6)/3)*(eq1.@ssr-eq5.@ssr)/eq5.@ssr
   simres(!rep,8)=eq1.@r2
next
simres.write f:\MCS\mcs43sim.txt
workfile f:\MCS\mcs43sim.wf1 u 1 !R
read f:\MCS\mcs43sim.txt dw BP1 LMt1 LM2 LMF2 LM3 LMF3 R2
genr rejBP1=BP1>@qchisq(0.95,1)
genr rejLMt1L=LMt1<@qtdist(0.05,!n-4)
genr rejLMt1R=LMt1>@qtdist(0.95,!n-4)
genr rejLMt1=abs(LMt1)>@qnorm(0.975)
genr rejLM2=LM2>@qchisq(0.95,2)
```

Table 4.5. Rejection probabilities of serial correlation tests, $n = 30, \rho = 0.4$.

	BP1	LM2	LM3	LMF2	LMF3	LMT1	LMT1L0	LMT1R
$\overrightarrow{p}$	0.283	0.223	0.180	0.194	0.156	0.268	0.003	0.353
$\overrightarrow{\mathrm{SD}}(\overrightarrow{p})$	0.005	0.004	0.004	0.004	0.004	0.004	0.001	0.005

```
genr rejLMF2=LMF2>@qfdist(0.95,2,!n-5)
genr rejLM3=LM3>@qchisq(0.95,3)
genr rejLMF3=LMF3>@qfdist(0.95,3,!n-6)
```

Table 4.5 shows that for $n = 30$ the best power is obtained for testing in the direction of the true DGP, that is, for positive first-order serial correlation, and that we have to face decreasing power when we test two-sided, or for second or third order serial correlation too. Note that for a moderate n the type II error probabilities are still quite substantial. Although the tests that use the χ^2 critical values show higher rejection probabilities than their F-test counterparts this should not be interpreted as higher power, because we have already established that they also reject a true null hypothesis with higher probability. Any genuine power difference can only be established after size correction. However, because the modification of the crude asymptotic test just involves a multiplication of the test statistic by the fixed number $(n - k - l)/(nl)$, where l is the order of serial correlation tested for and k the number of regressors in the original model, it is self-evident that after proper size correction the two test procedures will have equivalent power.

Of course, the rejection probabilities will improve for larger values of n and for larger values of ρ, the serial correlation coefficient. For $n = 200$ and $\rho = 0.4$ one finds that all estimated rejection probabilities are larger than 0.99 apart from that of LMT1L which is 0.000.

4.4 Results for a Model with Jointly Dependent Variables

We continue by analyzing OLS and Two-Stage Least-Squares (TSLS) in a simultaneous model by program mcs42.prg. It is about a simple two equation simultaneous system

$$\left.\begin{aligned} C_t &= \beta_1 + \beta_2 Y_t + \varepsilon_t \\ Y_t &= C_t + G_t + I_t \end{aligned}\right\}, \tag{4.19}$$

where $\varepsilon_t \sim NIID(0,\sigma^2)$ and the variables G_t and I_t are exogenous. It is a stylized version of a macro model in which C_t is consumption, Y_t income, G_t government spending and I_t investments in year t. We generate the series in such a way that I and G are not uncorrelated. Moreover, they are "constant in repeated samples," because we will try to make a fair comparison between results in a small and in a large sample. So, for $t = 1, \ldots, T$ we have (// indicates integer division, which is division with neglect of the remainder)

$$\begin{aligned} I_t &= 20 + 0.5[t - 10 * (t-1)//10] \\ G_t &= 10 + 0.3[t - 10 * (t-1)//10] - 0.02[t - 10 * (t-1)//10]^2. \end{aligned}$$

Hence, after every 10 observations the same observations are obtained again. Therefore, the $T \times 3$ matrix Z with $z_t' = (1,\, I_t,\, G_t)$ obeys

$$\lim_{T\to\infty} \frac{1}{T} Z'Z = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} z_t z_t' = \frac{1}{10} \sum_{t=1}^{10} z_t z_t' = \frac{1}{20} \sum_{t=1}^{20} z_t z_t'.$$

In the program we use CC for C because the latter refers to the intercept. The data for Y are generated via the reduced form for Y.[3] The consumption function is estimated both by OLS and TSLS and the latter results are used for testing the value of β_2. We compare results for $T = 20$ and $T = 1{,}000$ and we will discuss $T \to \infty$. For $T = 20$ in eq5 the reduced form for Y is estimated and its residuals are added in eq6 to the structural equation for CC. By this auxiliary regression we can perform the Durbin–Wu–Hausman test for endogeneity of Y.

```
'mcs44.prg MCS of IV
workfile f:\MCS\mcs44.wf1 u 1 1000
genr t=@trend(0)               'obs 1 .. 1000
genr tt=t-10*@floor((t-1)/10)  'obs 1..10,1..10,..
genr I=20+0.5*tt               'proxy for Investments
genr G=10+0.3*tt-0.02*tt^2     'proxy for Government spending
!beta1=50             'intercept Consumption equation
!beta2=0.8            'income coefficient in consumption equation
!sigma=3              'disturbance standard deviation
rndseed 9876543210
```

[3] Note that it is technically impossible to generate both the endogenous variables via the structural form. Hence, we face the following paradox: Economic agents do have qualities that computers lack.

```
!R=10000
matrix (!R,7) simres
for !rep=1 to !R
   smpl @all     'all 1000 observations are used
   genr eps=!sigma*nrnd               'iid Normal disturbances (0, sigma^2)
   genr Y=(!beta1+G+I+eps)/(1-!beta2) 'Income Y according to reduced form
   genr CC=!beta1+!beta2*Y+eps      'Consumption CC according structural form
   equation eq1.ls CC C Y           'OLS Consumption function
   simres(!rep,1)=eq1.@coefs(2)
   equation eq2.tsls CC C Y @ C I G  'TSLS Cons function, instruments: C I G
   simres(!rep,2)=eq2.@coefs(2)
   simres(!rep,3)=(eq2.@coefs(2)-!beta2)/eq2.@stderrs(2)
   smpl 1 20                        'just first 20 observations
   equation eq3.ls CC C Y           'OLS Consumption function
   simres(!rep,4)=eq3.@coefs(2)
   equation eq4.tsls CC C Y @ C I G
   equation eq5.ls Y C I G          'OLS reduced form
   eq5.makeresids res_eq5           'residuals eq5
   equation eq6.ls CC C Y res_eq5  'auxiliary regr. Durbin-Wu-Hausman test
   simres(!rep,5)=eq4.@coefs(2)
   simres(!rep,6)=(eq4.@coefs(2)-!beta2)/eq4.@stderrs(2)
   simres(!rep,7)=eq6.@tstats(3)   'DWH test staitsic
next
simres.write f:\MCS\mcs44sim.txt
workfile f:\MCS\mcs44sim.wf1 u 1 !R
read f:\MCS\mcs44sim.txt b2 b2iv tb2iv b2s b2ivs tb2ivs DWH
genr rejectDWH= abs(DWH)>@qnorm(0.95)
genr rejecttb2iv= tb2iv>@qnorm(0.95)
genr rejecttb2ivs= tb2ivs>@qnorm(0.95)
```

After running the program we find in workfile mcs44.wf1 the results from the final replication. From these we collect some figures in Table 4.6.

Table 4.6 shows that in this particular arbitrary single replication for both sample sizes the results for $\hat{\beta}_2$ and $\hat{\sigma}$ are much closer to their true values for the consistent TSLS estimates than for the inconsistent OLS, whereas the $SE(\hat{\beta}_2)$ values are (as is usual) much smaller for the larger

Table 4.6. Results final replication of mcs44.prg.

		$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\sigma}$	SSR
$n = 20$	OLS	0.9221	0.0179	1.232	27.33
	TSLS	0.8480	0.0350	1.720	53.27
$n = 1{,}000$	OLS	0.9485	0.0026	1.420	2012.51
	TSLS	0.8164	0.0093	2.681	7173.70

sample size, but also inappropriately suggest a much better precision of the OLS estimates than for the TSLS estimates. This has two reasons. It is partly due to the much smaller SSR value for OLS, which leads to a smaller σ estimate. That SSR is smaller for OLS than for TSLS follows directly from the mere fact that OLS minimizes the sum of squared residuals, whereas TSLS does so too, but not for these residuals, but for those from the so-called second stage regression. The other reason we will not discuss in full detail. We just explain what would have occurred if there were only one regressor x_i and one instrument z_i. Comparing the then relevant expressions $\sigma^2/\Sigma x_i^2$ and $\sigma^2/[\Sigma x_i z_i(\Sigma z_i^2)^{-1}\Sigma x_i z_i]$ by taking their ratio, we find $[\Sigma x_i z_i(\Sigma z_i^2)^{-1}\Sigma x_i z_i]/\Sigma x_i^2 = \rho_{xz}^2 < 1$. So, that $SE(\hat{\beta}_2)$ is smaller for OLS than for IV is simply due to algebra, and their difference will be more severe the weaker the instruments are. This difference does not have a statistical reason, because the OLS coefficient estimator and the estimator of its variance are both inconsistent. Therefore these do not carry information on β_2 and on the OLS estimation error that can be interpreted in the usual way.

In the present case, with more regressors and instruments than just one, the weakness or strength of the instruments can be expressed by statistics from the first-stage regression (eq5 in the program). The F-statistic of 8.93 indicates that the instruments are quite weak when $n = 20$, but running the same regression for $n = 1{,}000$ yields 194.28 which can be classified as reasonably strong. Note that because the population correlation between Y_i and ε_i is one, the simultaneity is extremely severe. At both sample sizes the OLS results look terrific, with $R^2 = 0.99$, though even at $n = 20$ an investigation of the simultaneity by the Durbin-Wu-Hausman test (which has p-value 0.0000) already indicates that Y is not exogenous. From workfile mcs44sim.wf1 we can investigate the actual distribution of the relevant OLS and TSLS statistics.

The results collected in Figure 4.3 illustrate the systematic bias in the OLS results, which is not resolved by increasing n. The degree of overidentification is one for the consumption equation. Therefore, the expectation of the TSLS estimator exists, but not its higher-order moments. This clearly has practical consequences at $n = 20$ but not at $n = 1{,}000$. The TSLS coefficient estimator is remarkably well centered

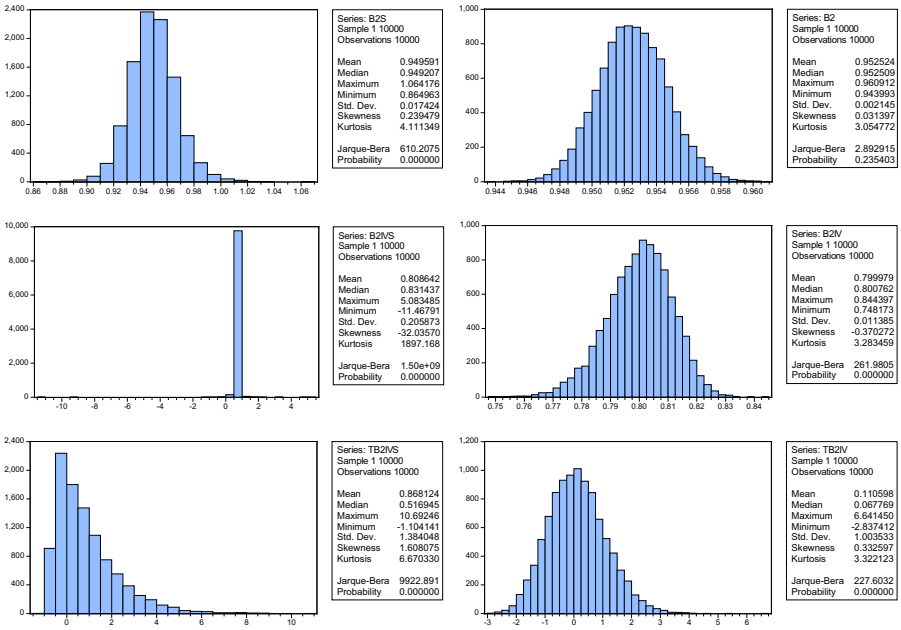

Fig. 4.3 OLS and TSLS for $n = 20, 1{,}000$ examined by mcs44.prg.

around the true value, but its convergence to normality is rather slow. Especially at the small sample size the usual coefficient test procedure will therefore be inflicted with problems for controlling the size of tests.

4.5 Results for Nonlinear Models for a Binary Variable

In program msc45.prg we examine and compare probit and logit estimation in a simple model for a binary dependent variable. Apart from the intercept there are two explanatory variables, which are drawn from the standard normal but in such a way that they are correlated (if $rho23 \neq 0$ in the program). Once they are drawn they are kept fixed throughout the replications. For the latent variable y_i^* and the observed variable y_i in the probit model we have

$$y_i^* = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + u_i, \quad u_i \sim NIID(0,1),$$
$$y_i = \mathbb{I}(y_i^* > 0), \quad i = 1, \ldots, n.$$

The deterministic part of y_i^* is called the index. In the logit model the disturbances are drawn from the logistic distribution,[4] which has variance $\pi^2/3$. To make the probit and logit results easily comparable we multiply the coefficients in the logit model by $\pi/\sqrt{3}$. Note that given the chosen nature of the explanatory variables the index is on average equal to β_1. Therefore, β_1 is the major determining factor of the frequency of the occurrence of zeros and ones for y_i in a sample of size n. If $\beta_1 = 0$ this will be close to 50–50. For β_1 larger in absolute value than 2, depending on the size of n, a sample may frequently just contain zeros, or just ones. Then maximum likelihood estimation is impossible. In fact, also when the fraction of zeros or ones is rather small/large, estimation may be impossible due to the so-called perfect classifier problem. When this occurs in a particular replication the program will stop and mention "overflow." Nonlinear models are estimated via an iterative procedure, in which estimates are improved step by step until the extremum of a criterion function has been found (according to a convergence criterion). The iterative process has to start from well-chosen initial values for the parameters. To speed up the process in the simulation we provide by the param statement again the true values of the parameters.

```
'msc45.prg  MCS of probit and logit models
!n=100
workfile f:\MCS\mcs45.wf1 u 1 !n
rndseed 12345
genr x2=nrnd
!rho23=0.5
genr x3=!rho23*x2+@sqrt(1-!rho23^2)*nrnd
!beta1=1.5
!beta2=1
!beta3=1
!resc=3.14159/@sqrt(3)
!beta1rs=!beta1*!resc
!beta2rs=!beta2*!resc
!beta3rs=!beta3*!resc
genr index=!beta1+!beta2*x2+!beta3*x3   'X*beta
```

[4] In EViews @rlogistic directly yields drawings ζ with CDF $\eta = F(\zeta) = e^{\zeta}/(1 + e^{\zeta})$. These are in fact obtained by the inverse transformation method of Section 2.1, using $\zeta = F^{-1}(\eta) = \log[\eta/(1 - \eta)]$. Hence, after genr eta=rnd, they can also be obtained by genr zeta=log(eta/(1 − eta)).

```
!R=10000
rndseed 9876543210
matrix (!R,4) simres
for !rep=1 to !R
   genr ylatentp=index+nrnd     'index plus iid drawings from N(0,1)
   genr ysimp=(ylatentp>0)      'simulated 0-1 variable y
   param c(1) !beta1 c(2) !beta2 c(3) !beta3
   equation eqp.binary(d=p) ysimp c x2 x3 'd=p means probit estimation
   simres(!rep,1)=eqp.@coefs(2)
   simres(!rep,2)=eqp.@stderrs(2)
   genr ylatentl=index*!resc+@rlogistic
         'rescaled index plus iid drawings from logistic distribution
   genr ysiml=(ylatentl>0)            'simulated 0-1 variable y
   param c(1) !beta1rs c(2) !beta2rs c(3) !beta3rs
   equation eql.binary(d=l) ysiml c x2 x3 'd=l means logit estimation
   simres(!rep,3)=eql.@coefs(2)/!resc
   simres(!rep,4)=eql.@stderrs(2)/!resc
next
simres.write f:\MCS\mcs45sim.txt
workfile f:\MCS\mcs45sim.wf1 u 1 !R
read f:\MCS\mcs45sim.txt b2p seb2p b2l seb2l
genr t2p0=(b2p-!beta2)/seb2p
genr rejt2p0l=t2p0<@qnorm(0.05)
genr rejt2p0r=t2p0>@qnorm(0.95)
genr t2l0=(b2l-!beta2)/seb2l
genr rejt2l0l=t2l0<@qnorm(0.05)
genr rejt2l0r=t2l0>@qnorm(0.95)
```

From Figure 4.4 it is obvious that at $n = 100$ the asymptotic normality of the maximum likelihood estimators and of the null distribution of the tests is not eminent yet. Also note the bias of at least 10% in the coefficient estimates. Nevertheless, the consequences for test size are not really dramatic, as can be learned from Table 4.7.

A warning is in place that the few results just obtained on logit and probit analysis do not allow to draw general conclusions on the accuracy of inference in such models. After all, we only examined models with two particular stylized explanatory variables, just for $n = 100$, and investigated one combination of specific values for the parameters ρ_{23}, β_1, β_2, and β_3 only. This highlights an undeniable limitation of MCS: its specificity. How should the experiments be designed in order to be able to deduce reasonably general recommendations? This and various other more or less related issues will be discussed in the next section.

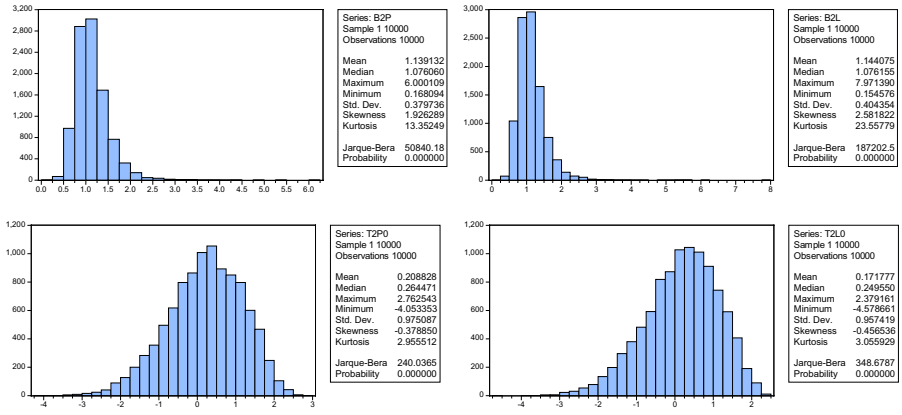

Fig. 4.4 Probit and Logit for $n = 100$ examined by mcs45.prg.

Table 4.7. Rejection probabilities of coefficient tests for probit and logit, $n = 100$.

	probit-L	probit-R	logit-L	logit-R
$\overrightarrow{p}$	0.0385	0.0566	0.0411	0.0442
$\overrightarrow{\mathrm{SD}}(\overrightarrow{p})$	0.0019	0.0023	0.0020	0.0021

Exercises

1. What value of R would roughly be required to establish the size corrected critical values in Section 4.2.1 with a standard error not exceeding 0.01?
2. Establish size corrected 2.5% one-sided critical values for TB2 and their standard errors for the cases of both Figures 3.5 and 3.6.
3. Adapt the RESET program mcs41.prg and examine for the various test procedures the seriousness of loss of control over type I error probability when having nonnormal disturbances with a skew distribution. Run your programs with a moderate and with a large value of R and examine how this affects your results (realizing that this cannot have any effect on the actual properties of the RESET test that are being investigated, but only on the quality of MCS analysis of those properties).

4. Adapt program mcs41.prg such that you examine the type I and II error probabilities when you change the factor 0.5 of the sinus contribution to x in 0.7 and explain the changes.
5. When running mcs42.prg for $n = 30$ we obtained an average R^2 (coefficient of determination) of 0.983 and for $n = 200$ we obtained an average R^2 of 0.994. Both are very high. Due to the stability of the model and the stationarity of x_t one would expect that the R^2 is in principle invariant with respect to n. The different values can be explained as follows: (a) the warming-up period of 100 is too short; (b) the differences are just a result of the MCS randomness of these average R^2 values; (c) the x_t series is kept fixed over the replications and hence may have different characteristics over the first 30 observations in comparison to the 200 observations. Try to find out which of these three explanations are true and do matter.
6. Use program mcs42.prg to examine the dependence of the size problems of the tests on the value of σ.
7. Use program mcs42.prg to examine the dependence of the size problems of the tests on the value of β_2. Try $\beta_2 = 1$ and $\beta_2 = .6$.
8. Adapt program mcs42.prg such that the DGP is still ARX(1) but that serial correlation tests are being employed in the static model, which does include the intercept and x_i but omits y_{i-1}. What are the probabilities to detect this form of misspecification by serial correlation tests for small and substantial n, and how does this compare to testing directly for the omission of y_{i-1}? Can the Durbin–Watson test here be used safely?
9. Use program mcs43.prg to examine the dependence of the power of the tests on the values of σ and β_2.
10. Run program mcs44.prg and use its results to examine the type I error probabilities of tests for β_2 and draw your conclusions. Note that it is worthwhile to also examine the tests against one-sided alternatives. Do so (without adapting and

re-running the program, but simply by generating a few further variables in workfile mcs44sim.prg).

11. Examine the rejection probability of the DWH test and comment. Note that the impressive result would not mean very much if a true null hypothesis would also be rejected with high probability.
12. Increase and decrease the value of σ in program mcs44.prg and try to understand its effects on the fit of the regressions, on the strength of the instruments, on the seriousness of the simultaneity, on the distribution of OLS and TSLS coefficient estimators and tests.
13. In program mcs45.prg try different values of β_1 and of ρ_{23}, such as $\beta_2 = 0, 2$ and $\rho_{23} = 0, .8$. Are some of these values problematic regarding estimation, estimator bias, size of the tests? Are they still when you have $n = 500$?
14. Adapt program mcs45.prg such that the data generated with the normal disturbances are analyzed by logit and those with logistic disturbances are analyzed by probit and investigate how seriously such a misspecification of the actual disturbances affects inference.

5

Further Issues Regarding Classic MCS

In the foregoing we explored the basic purposes and means of naive or classic Monte Carlo simulation studies. These are to produce estimates of particular functions of random variables by experimentation. These functions mostly involve integrals too complex to be evaluated analytically. They represent moments or a CDF of the random elements of a vector or matrix. The scalar random variable of interest, denoted q_n, may represent an estimator of a population characteristic, possibly a parameter or perhaps a variance, or it is a function of various estimators to be used as a test statistic. Our primary interest lies in scalar random variables q_n which are continuously distributed with density $f_{q_n}(q;\theta)$, where θ is the population parameter vector and n the size of the sample from which q_n is obtained. Often the integrals of interest cannot be evaluated analytically because an explicit expression for $f_{q_n}(q;\theta)$ is simply not available. Therefore we use MCS experiments to generate approximations to these integrals, which usually remain implicit. However, when establishing a moment it of course concerns $E(q_n) = \int q f_{q_n}(q;\theta)dq$, and when assessing a probability we evaluate a CDF such as $F_{q_n}(Q) \equiv \Pr(q_n < Q) = \int_{-\infty}^{Q} f_{q_n}(q;\theta)dq$, and

in case of quantiles we seek a solution Q_p to the integral-equation $F_{q_n}(Q_p) = \int_{-\infty}^{Q_p} f_{q_n}(q;\theta)dq = p$, for given $0 < p < 1$.

As we saw in the preceding sections, on a computer one can draw a great number of realizations of q_n from which MCS estimates can be constructed for the moments (if these exist), and probabilities or quantiles of interest, provided the DGP is fully specified. These MCS estimates converge for increasing number of MCS replications to their true values and their accuracy for a finite but substantial number of replications can be controlled by using well-known results from mathematical statistics.

A major limitation and complication of designing a useful study based on classic Monte Carlo simulation is the requirement to fully specify the DGP. Within a particular class of models a DGP specifies numerical values for the observations on all the involved deterministic explanatory variables, a full parametrization of the distribution of all its random variables and the actual parameter values, such that realizations of all random endogenous variables involved can be generated on a computer. Of course, one does not have to restrict the MCS study to examining just one particular DGP, as we did most of the time in the examples of the preceding sections. One can consider a class or family of models and run simulations over a discrete grid of numerical values for its parameters and alternative choices of the deterministic variables. However, when one wants to produce rather general recommendations regarding the qualities of an estimation technique or test procedure many crucial choices will have to be made regarding the scope of the simulation design. In making all these choices (which sample sizes, what number and type and realization of exogenous regressors, what distribution for the disturbances, what specification of econometric model, what particular grid of parameter values, etc.) one is explicitly or implicitly using some kind of MCS methodology. In what follows we will try to address many aspects of such a methodology, leading to a list of ten aspirations and recommendations. But, before we get to that, we will first mention a few landmarks in the history of the development of MCS, and also focus on various not yet discussed aspects of designing and delivering a MCS study which aim to enhance its quality.

5.1 Some Landmarks in the History of MCS

Games of chance have always plaid a major role in the development of probability theory and mathematical statistics. In the 18th and 19th century important inventions such as crude forms of the LLN and CLT have been suggested by Bernoulli and Laplace, respectively, and like most probabilists they were fascinated by gambling. In present day Monte Carlo experiments, however, the inspiration for econometricians is not so much the tiny probability of the gambler of making a huge profit once in a lifetime, but the almost certainty of the owner of the casino to make sizeable gains almost every single day.

We give a few early examples of using random experimentation that paved the way for present day Monte Carlo simulation, see also Stigler (1999). In 1873 Asaph Hall reports on the experimental determination of the number π by throwing a needle repeatedly, and making measurements of certain angles. Such experiments were already described earlier by Georges Louis Leclerc Comte de Buffon (1707–1788), and are therefore known as the Buffon experiment.

Sir Francis Galton (1822–1911), who discovered in the course of the 1880's the concept of regression, was a very practicable gentleman scientist. In 1873 he invented a device that he called the quincunx by which a histogram of the normal distribution can be synthesized through experimentation by literally integrating a succession of binary outcomes. This experiment involves pouring out bullets onto needles placed in a particular pattern in a vertical slit, which is transparent at one side. It visualizes the CLT. He also designed a two-stage quincunx that demonstrated conditional probability. In 1890 he published on how he could generate drawings from the normal distribution, in fact the $N(0, 1.483^2)$ distribution for which the 3rd quartile is 1, by using three dice in an ingenious way. A role by one die allows in fact 24 different outcomes instead of six, when all four edges of the six sides are considered separately. One die determined the signs of the next few outcomes, one other die the outcomes in the center of the distribution or (indicated by outcome *, which could occur with probability 3/24) whether the third die should be rolled to determine an outcome in the tails. In this simple mechanical way all quantiles corresponding exactly

with probabilities $1/2 \pm (2i-1)/48$ for $i = 1, \ldots, 21$ and $1/2 \pm [3/8 + (2i-1)/(8 \times 48)]$ for $i = 1, \ldots, 24$ could be generated.

William Sealy Gosset (1876–1937), better known under his pseudonym Student, published in 1908 about the following MCS experiment, not performed on a computer of course, but by hand. He wrote on 3000 cards both the body length and the length of the middle finger of 3000 criminals, and calculated the population correlation coefficient. Next he estimated the correlation coefficient from a random sample of size $n = 4$ and he replicated this $R = 750$ times to check the formula that he had derived for the sampling distribution of the estimator of the correlation coefficient. He also used such experiments to build faith in the distribution of the statistic that he had derived and which we now know as one that follows a Student distribution.

The first article explicitly addressing such simulation techniques as explored here as "the Monte Carlo method" seems to be Metropolis and Ulam (1949). They indicate a variety of problems in the natural sciences, one of which being the evaluation of an integral, which can be solved with reasonable accuracy by preforming random experiments, possibly on a modern computing machine. They urge for the further refinement of methods that would allow to suit the size of the simulation sample to the desired accuracy.

Studies in econometrics that already have the major characteristics of present day MCS studies appeared since the 1950's. For instance, Wagner (1958) reports extensively on a Monte Carlo study of estimates and test statistics for simultaneous linear structural equations in samples of size $n = 20$, carried out on an IBM card programmed calculator on which he produced $R = 100$ replications. He examines in great detail the inaccuracies of the simulation approach.

5.2 Options to Enhance the Quality of MCS Studies

In this section we mention various approaches by which MCS studies can be improved in some way or another. We first discuss ways to make MCS analysis less naive, and gain efficiency by incorporating particular information which may either be directly available or can be derived by analytical methods. Next we mention and refer to various options that

can be used to improve the presentation of MCS results and which help to better understand and interpret their major findings. This is followed by some general remarks on conditioning in the context of MCS. Many actual studies pay little to no attention to this. Therefore we address major issues and their consequences with respect to conditioning in two further separate sections.

5.2.1 Reducing Simulation Variance by Exploiting Information

There are two general methods to enhance the accuracy and efficiency of the relatively simple MCS estimators discussed before. These allow to adapt the standard Monte Carlo estimators such that one can achieve the same accuracy as naive MCS with fewer replications. One of the procedures involves using so-called "antithetic variables" and the other exploits "control variates"; for a more complete treatment see Hendry (1984, 1995) and Davidson and MacKinnon (1993).

When using antithetic variables one deliberately generates the drawings $\{q_n^{(r)}; r = 1, \ldots, R\}$ such that they are no longer independent. For instance, let $q_n^{(r)}$ be determined by an $n \times 1$ vector of random disturbances $u^{(r)}$, where the distribution of the individual elements of $u^{(r)}$ is symmetric around zero, and let R be even. Then one can generate $q_n^{(2l-1)}$ for $l = 1, \ldots, \frac{R}{2}$ using disturbance vector $u^{(2l-1)}$ in the odd replications and generate $q_n^{(2l)}$ for $l = 1, \ldots, \frac{R}{2}$ using the same disturbances in the even replications after swapping their sign, so taking $u^{(2l)} = -u^{(2l-1)}$. Not only is this cheaper because fewer random numbers have to be generated. It also assures symmetry around zero in the MCS sample of the elements of the disturbance vectors, which perfectly matches its population distribution. Moreover, it leads to a smaller variance of $\overrightarrow{E}(q_n)$, when $q_n^{(2l)}$ and $q_n^{(2l-1)}$ are negatively correlated.

When using control variates one incorporates known analytical results in the MCS study and does not attempt to establish for instance $E(q_n)$ from scratch, but uses another random variable, say $\tilde{q}_n$, which is correlated with q_n and for which particular analytical results on its distribution are available, for instance $E(\tilde{q}_n)$. Then the discrepancy between $\overrightarrow{E}(\tilde{q}_n)$ and $E(\tilde{q}_n)$ can be used to improve upon $\overrightarrow{E}(q_n)$ as an approximation of $E(q_n)$.

Although both these techniques, which enable to reduce computer labor by exploiting brain power, are intellectually appealing, they become less of a necessity in an era with almost unlimited computer force. Therefore we do not pay further attention to them here.

5.2.2 Presenting Findings Transparently

When a MCS study focusses on one particular inference technique for just one particular DGP presenting the numerical results can be quite straightforward because it just concerns a few numbers and an indication of their accuracy. However, often the study aims at greater generality covering classes of models over substantial ranges of their multidimensional parameter space and comparing various alternative inference techniques. Then presenting the numerical results from MCS experiments in such a way that this supports drawing the appropriate conclusions is certainly not self-evident.

A rather sophisticated form of summarizing the results of MCS experiments produces a so-called response surface. These aim to achieve results similar to those from analytical derivations based on asymptotic expansions such that for various n and any values of the nuisance parameters the bias, efficiency or rejection probabilities are represented in a formula. The formula is obtained by using regression techniques to fit a surface to the scatter of MCS estimates over the grid of parameter values and sample sizes examined. Hence, in these regressions the different values in the experiments for the nuisance parameters and for the sample size (n) form the basic explanatory variables. Because these surfaces usually have to be highly nonlinear (bias and the other phenomena of interest depend usually nonlinearly on the nuisance parameters and sample size) finding an appropriate specification of the functional form is usually nontrivial (unless the analytical results are already available). Therefore the hazard of misspecification is serious. In the past the interpolations and extrapolations suggested by response surfaces have proved to be seriously misleading occasionally.

Collecting all numerical estimates for all individual DGP's examined in a great number of tables is one of the least attractive forms of presenting the results from an extensive MCS study. As a rule a deliberate and well balanced selection should be made, in which the

major results should be grouped and discussed in such a way that the worthwhile findings are adequately and convincingly illustrated. Representing the numerical results graphically is usually much more effective. Useful ways of graphical representation of rejection probabilities of tests are p-value plots and size-power curves, see Davidson and MacKinnon (1998). A limitation of plots is that paper has only two dimensions. Of course one can collect more graphs in one figure to represent phenomena which are determined by more than just one factor, or one can project a 3D surface on paper, provided it still allows the observer a proper perspective. Using various colors offers another opportunity to put more information in just two dimensions, whereas on screen one can display animated 2D or 3D pictures. Hence, it is possible to represent phenomena in more than 3 dimensions graphically.[1]

5.2.3 Clarity About Conditioning

In the EViews programmes used for illustrations hence far, the regression models usually contained one or more exogenous regressors. Occasionally these were constructed rather mechanically (as in programs mcs22, mcs32, mcs41, mcs44), but in other cases these were realizations of IID random variables (in programs mcs13, mcs34, mcs35, mcs45) or of autoregressive stationary processes (programs mcs14, mcs42, mcs43) or a nonstationary process (program mcs33). However, when we used a randomly drawn exogenous regressor we did not generate a new random realization in each replication. Hence, in fact we conditioned the analysis on that particular perhaps accidently rather atypical random realization of a particular distribution. Therefore, we did not necessarily obtain results representative for the type of process chosen for the exogenous regressor. So, would it not have been better if we had redrawn random exogenous series in every replication?

The major issue seems to be here: what does one want to learn from the simulation? If the primary interest lies in examining the finite sample behavior of particular inference techniques in a model for particular empirical data involving specific exogenous variables, it would seem

[1] Examples of that can be seen at http://www.feb.uva.nl/ke/jfk.htm by surfing to animated (flash player required) diagrams of Kiviet and Niemczyk (2011).

best to use those actual realizations of the exogenous variables in all replications. Then the results will be highly specific and probably just valid for that particular situation, but that should not really matter. The very specific question receives a very specific (conditional) but therefore highly appropriate answer.

The situation is different if one wants an answer to a more general question, such as: What is the actual significance level of a particular test procedure in a certain class of models when the exogenous regressors have the characteristics of a random walk and the sample size is some particular value n. Then it seems essential to renew the generation of the exogenous random walk regressor in every replication. Conditioning the whole Monte Carlo study on just one random realization of the random walk process would still yield an unbiased estimator of the significance level, but not a large-R consistent one. Hence, irrespective of the magnitude of R, a component of the Monte Carlo estimation errors will be nonnegligable and not under control either, because its variance cannot be estimated when we condition on just one realization. These issues are addressed more formally in Edgerton (1996). However, he concludes that one should never condition on just one realization of a random exogenous regressor. We don't agree. Indeed, one should not condition if one wants to analyze characteristics of the unconditional distribution. But, a conditional distribution may be of interest in its own right. Moreover, one could analyze both the unconditional distribution and the conditional one, the latter possibly for a few separate realizations. After all, the information obtained by an unconditional MCS is not all that useful if one has to employ it in a situation where the actual exogenous variables have been observed already, as in many time-series applications. Unconditional results may be more appropriate in a cross-section and a microeconometric panel context, especially when the sample has been drawn from a large population and is to be used for inferences on that population. In time-series econometrics the notion of a population is usually less apposite. There, the DGP, entailing the specification of a population distribution of the random error terms, is used as its substitute. Thus, in that context the whole issue simply hinges upon whether in the specification of the DGP any exogenous variables are either fixed or random, and if they are random,

whether the MCS study should inform us on the qualities of inference conditioned (or not) on a specific realization.

Clearly, MCS results obtained by conditioning on a specific process for an exogenous variable drawn randomly from a family of processes should certainly not be used to provide answers to questions regarding the whole family. On the other hand, MCS results obtained for whole families seem inappropriate to provide specific answers about particular family members too. For instance, it could be the case that particular individual family members give rise to serious overrejection of a test, and other members to underrejection, whereas an unconditional analysis would not detect any problems. In practice, if one has the actual observations on the exogenous variables of the model available, then it seems best to condition MCS as much as possible on all the information at hand. As we will see later, this is one of the attractions of both dedicated simulation studies and of bootstrap simulations, in which one focusses as much as possible on the typical characteristics of one actual empirical process.

Of course, investigators should always mention whether they kept randomly drawn exogenous regressors fixed or not over the replications, but not that many do.

5.3 Conditioning on Exogenous Variables

Here we shall illustrate the effects that conditioning on exogenous variables may have by making comparisons with unconditional results, where exogenous variables are redrawn every replication. We will focus on a very specific but generic case, which demonstrates that conditioning or not may have profound effects. In the course of the analysis, as a side line, but with much broader relevance than its appearance in the present context of conditioning issues, particular fundamental issues arise regarding the desirability to transform the parameter space of the DGP into a lower dimensional MCS design parameter space. The former simply refers to the parameters and their domain as they appear in the formulation of the DGP, whereas the latter may have some parameters in common with the DGP, but preferably has all its parameters referring to particular autonomous econometric notions. Therefore, it entails

particular desirable orthogonality properties. Moreover, it may have lower dimensionality than the DGP parameter space, because it has absorbed invariance properties when these could be derived analytically. Introducing and discussing these parametrization issues in the context of a simple actual example makes it easier to incorporate them later in this section in a more general framework when we develop a comprehensive methodology for setting up MCS investigations.

5.3.1 Orthogonalizing the Parameter Space

In the present illustration we focus on a very basic partial static simultaneous equations model. Let u_i, z_i, and w_i be mutually independent zero mean $NIID$ drawings for $i = 1, \ldots, n$ with standard deviations $\sigma_u > 0$, $\sigma_z > 0$, and $\sigma_w \geq 0$, respectively. Our parameter of interest is β_2 in the structural equation

$$y_i = \beta_1 + \beta_2 x_i + u_i, \tag{5.1}$$

where x_i is given by the reduced form equation

$$x_i = \pi_1 + \pi_2 z_i + v_i, \tag{5.2}$$

with reduced form disturbance

$$v_i = \phi u_i + w_i. \tag{5.3}$$

Note that $E(z_i u_i) = 0$. This makes z_i an appropriate instrumental variable, provided the identification restriction $\pi_2 \neq 0$ holds.

For particular second moments we find

$$\begin{aligned} \sigma_{zx} &= \rho_{zx}\sigma_z\sigma_x = \pi_2\sigma_z^2 \\ \sigma_{vu} &= \sigma_{xu} = \rho_{xu}\sigma_x\sigma_u = \phi\sigma_u^2 \\ \sigma_x^2 &= \pi_2^2\sigma_z^2 + \phi^2\sigma_u^2 + \sigma_w^2, \end{aligned}$$

from which it follows that

$$\pi_2 = \rho_{zx}\sigma_x/\sigma_z, \quad \phi = \rho_{xu}\sigma_x/\sigma_u, \quad \sigma_w^2 = (1 - \rho_{zx}^2 - \rho_{xu}^2)\sigma_x^2.$$

From this last relationship and $\sigma_w^2 \geq 0$ it follows that the parameters ρ_{zx} (instrument strength) and ρ_{xu} (simultaneity) should have values within or on the circle with radius unity

$$\rho_{zx}^2 + \rho_{xu}^2 \leq 1. \tag{5.4}$$

We will focus on assessment by MCS of the probability density of the estimation error of the IV slope estimator

$$\hat{\beta}_2 - \beta_2 = \sum_{i=1}^{n}(z_i - \bar{z})u_i \Big/ \sum_{i=1}^{n}(z_i - \bar{z})x_i. \tag{5.5}$$

It is obvious that the right-hand side is invariant with respect to β_1, β_2, π_1 and σ_z, so without loss of generality we may choose $\beta_1 = \beta_2 = \pi_1 = 0$ and $\sigma_z = 1$. Moreover, σ_u/σ_x is a scale factor of $\hat{\beta}_2 - \beta_2$, which allows to choose $\sigma_u = 1$ and next interpret the chosen standard deviation of x_i as σ_x/σ_u, which has a close relationship with the signal to noise ratio of the equation of interest. Next to the sample size n, the DGP defined above is found to have only three basic design parameters, namely σ_x/σ_u and the couple ρ_{zx}, ρ_{xu}, which should obey (5.4). Simulating the density of $\hat{\beta}_2 - \beta_2$ for different values of these autonomous characteristics of the DGP is straight-forward.

Note that initially it seemed that the DGP would require to specify the values of eight parameters, namely: β_1, β_2, σ_u, π_1, π_2, σ_z, ϕ, and σ_w. But due to the discovered invariance properties choosing values of π_2, ϕ and σ_w suffices. However, directly choosing empirically relevant values for these three DGP parameters is not so obvious. This becomes easier through the three nonlinear parameter transformations

$$\pi_2 = \rho_{zx}\sigma_x, \quad \phi = \rho_{xu}\sigma_x, \quad \sigma_w = \sigma_x(1 - \rho_{zx}^2 - \rho_{xu}^2)^{1/2},$$

because ρ_{zx}, ρ_{xu}, and σ_x/σ_u directly refer to well-understood fundamental econometric notions. Also for the interpretation of the MCS results, yet to be obtained, it seems more relevant to learn how the IV estimation errors are affected by independent changes in the simultaneity severity, the instrument strength or the signal-noise ratio than with respect to π_2, ϕ, and σ_w. Nevertheless, we require values for the latter when programming, because, unlike the design parameters ρ_{zx}, ρ_{xu}, and σ_x/σ_u, they appear in the DGP formulae.

This parameter transformation will not only serve the interpretation of the simulation results, but also the process of choosing a design for the grid of parameter values to be examined. This grid should be effective and efficient in the sense that within the chosen computational limitations it should generate a maximum of information on the

phenomena under study. Imagine that we have decided that we want (can afford) to examine each of the three parameters in seven different values, giving $7^3 = 343$ parameter value combinations. When using the DGP parametrization we might select, for instance, and for better or for worse, all combinations from

$$\begin{aligned} \pi_2 &\in \{-4, -2, -1, 0, 1, 2, 4\} \\ \phi &\in \{-2, -1, -.5, 0, .5, 1, 2\} \\ \sigma_w &\in \{0, 1, 2, 4, 8, 16\}. \end{aligned}$$

Note that this implies a range of values (many more than 7) for each of σ_x, ρ_{xu}, and ρ_{zx}. These values are rather unevenly spread over their domain, whereas not all their combinations are included. Because the parameter space $\{\rho_{zx}, \rho_{xu}, \sigma_x/\sigma_u\}$ refers to three distinct autonomous econometric notions we will call it an orthogonal parametrization, and it seems obvious that it is much more suitable as a basis for constructing a grid of interesting values than using directly the DGP parameters for a base. If we start off by choosing values for ρ_{zx}, ρ_{xu}, and σ_x/σ_u, we immediately realize that taking $\sigma_x/\sigma_u = 1$ suffices, because the estimation errors are a multiple of σ_u/σ_x so the results from $\sigma_x/\sigma_u = 1$ can simply be rescaled to obtain results for any value of σ_x/σ_u. Moreover, it is also easily seen that the estimation errors will be invariant with respect to the sign of ρ_{zx}, and will simply be their mirror image when changing the sign of ρ_{xu}, so by running much fewer than 343 parameter value combinations, for instance

$$\begin{aligned} \rho_{zx} &\in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\} \\ \rho_{xu} &\in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}, \end{aligned}$$

which are 120, we would learn much and much more.

5.3.2 Effects of Conditioning on IV Estimation

Both variables z_i and w_i are exogenous with respect to u_i and one might choose to keep them fixed over the replications. Of course, that would lead to simulation results that are very specific for just those arbitrary realizations. We will examine here the differences in the results between fixing just z_i, fixing both z_i and w_i, and treating both as fully random.

Note that in a practical situation z_i would be observed, but w_i not. Whether this should have implications for whether they should be kept fixed or not will be discussed later.

In program mcs51.prg below the arbitrariness of single random draws of z_i and w_i has been mitigated a bit by modifying them such that their first two sample moments correspond to their assumed population moments, which are: both have zero mean, a zero correlation, and variance 1 and $(1 - \rho_{zx}^2 - \rho_{xu}^2)\sigma_x^2/\sigma_u^2$, respectively.

```
'mcs51.prg:   (un)conditional IV in a simple model
!n=50
workfile f:\MCS\mcs51.wf1 u 1 !n
!rxu=0.6
!rxz=0.1
!sigw=@sqrt(1-!rxu^2-!rxz^2)
rndseed 12345
genr z=nrnd
genr zf=(z-@mean(z))/@stdev(z)
genr w=nrnd
equation eq0.ls w c zf
genr wf=!sigw*resid/eq0.@se
rndseed 9876543210
!R=10000
matrix (!R,3) simres
for !rep=1 to !R
   genr u=nrnd
   genr z=nrnd
   genr w=!sigw*nrnd
   genr x=!rxz*z+!rxu*u+w
   genr y=u
   genr xf=!rxz*zf+!rxu*u+w
   genr xff=!rxz*zf+!rxu*u+wf
   equation eq1.tsls y c x @ c z
   equation eq1f.tsls y c xf @ c zf
   equation eq1ff.tsls y c xff @ c zf
   simres(!rep,1)=eq1.@coefs(2)
```

```
   simres(!rep,2)=eq1f.@coefs(2)
   simres(!rep,3)=eq1ff.@coefs(2)
   for !j=1 to 3
      if simres(!rep,!j)>50 then simres(!rep,!j)=50 endif
      if simres(!rep,!j)<-50 then simres(!rep,!j)=-50 endif
   next
next
simres.write f:\MCS\mcs51sim.txt
workfile f:\MCS\mcs51sim.wf1 u 1 !R
read f:\MCS\mcs51sim.txt biv bivf bivff
```

We chose a weak, but not an extraordinary weak instrument ($\rho_{zx} = 0.1$), substantial simultaneity ($\rho_{xu} = 0.6$) and sample size $n = 50$. Figure 5.1 shows that in comparison to the unconditional case (BIV) there are only very moderate effects of conditioning on just the relatively weak instrument z (BIVF), but conditioning on both z and w (BIVFF) has a notable effect and makes the density bimodal. This is a well-know effect, extensively documented in the literature. Precisely the same DGP has been examined in Nelson and Startz (1990), which initiated the literature on the effects of weak instruments. They parametrized their DGP in a different way, apparently suitable for deriving some finite sample results, but less insightful from a practitioners point of view, because their formulation does not allow to write the reduced form for x explicitly. In their paper, and a range of follow-up papers, it has been taken for granted, without much discussion, that it is appropriate to condition on both z and w. We tend to a different opinion here. Multivariate extensions of this static model are typically applied in a cross-section framework, where the data are often a random representative sample from a much larger population. Inferences are meant to pertain to the population, and not just to the sample. In the frequentist's approach toward inference, one does envisage what coefficient estimates one could have obtained if one had drawn a different sample of disturbances from the same distribution. However, this should not just concern the disturbances u_i, but also the reduced form disturbances v_i, including their component w_i. Arguments to condition on the instruments z_i seem easier to rationalize, but here too one might

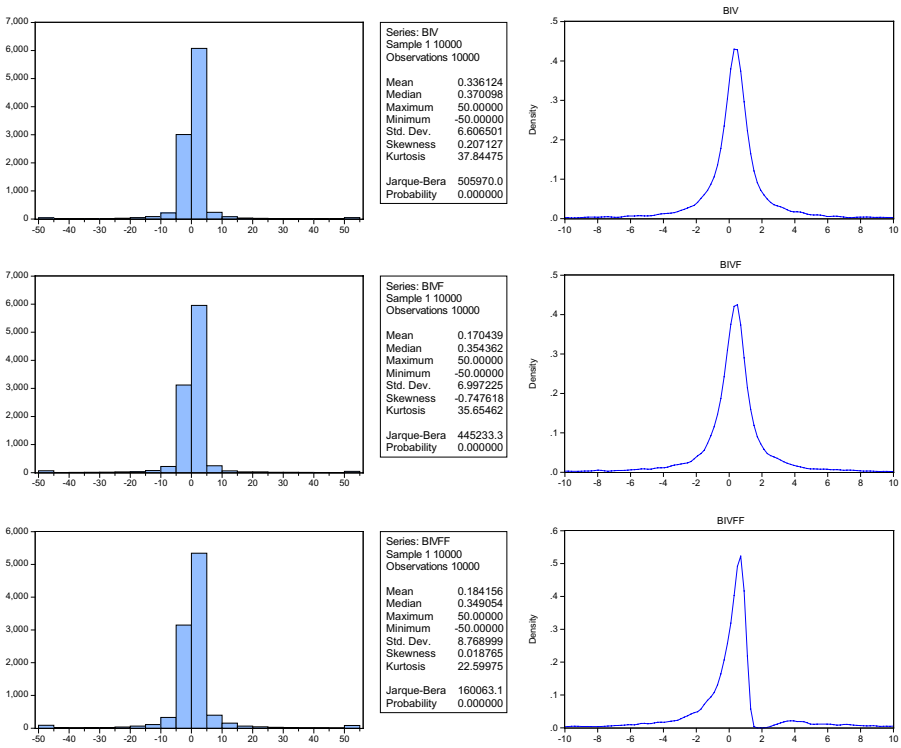

Fig. 5.1 Results from program mcs51.prg.

argue that if different individuals would have been sampled then different realizations of the instrumental variables would have been employed as well.

Note that by censoring in the program the IV estimates such that all are in the $[-50, 50]$ interval, we do not corrupt the estimate of their median, and (due to the nonexistence of moments) all other statistics mentioned next to the histogram have no simple interpretation anyhow, whereas the censoring is very beneficial for the quality of the kernel density estimates.

5.4 Conditioning on Endogenous Variables

Conditioning on endogenous variables may seem to be a very odd thing to do, but we already did for very good reasons in various of

the foregoing example programs, namely where we compared alternative estimators for the same parameter values (within programs mcs34, mcs35, and mcs44) or compared the size properties of alternative test procedures for the same null hypothesis (in both programs mcs42 and mcs43). All the statistics examined were calculated in the rth replication using the same disturbance vector and thus for the same dependent variables. Also when we compared rejection probabilities of a particular test procedure under a null with those under some alternative hypothesis (program mcs41) we did not estimate those two probabilities independently. This has particular consequences, mostly positive, which we will examine now.

Let $\hat{\beta}$ and $\hat{\beta}^*$ be alternative estimators for a scalar parameter β. Assume that we want to use MCS to find the bias of these two estimators and to learn in particular whether one is less biased than the other. If we estimate $\overrightarrow{E}(\hat{\beta})$ and $\overrightarrow{E}(\hat{\beta}^*)$ fully independently (not only between replications, but also within) then $\mathrm{Var}(\overrightarrow{E}(\hat{\beta}) - \overrightarrow{E}(\hat{\beta}^*)) = \mathrm{Var}(\overrightarrow{E}(\hat{\beta})) + Var(\overrightarrow{E}(\hat{\beta}^*))$ and next it is rather straight-forward to test whether $E(\hat{\beta}) - E(\hat{\beta}^*)$ is significantly different from zero and if it is, whether it should be rejected against the positive or the negative alternative. However, when the estimates $\hat{\beta}^{(r)}$ and $\hat{\beta}^{*(r)}$ are obtained in the same replication from the same random draws it is quite likely that they are positively correlated, which will lead to a smaller variance in the estimate of the difference in bias, since

$$\begin{aligned}&\mathrm{Var}(\overrightarrow{E}(\hat{\beta}) - \overrightarrow{E}(\hat{\beta}^*)) \\ &\quad = \mathrm{Var}(\overrightarrow{E}(\hat{\beta})) - 2\mathrm{Cov}(\overrightarrow{E}(\hat{\beta}), \overrightarrow{E}(\hat{\beta}^*)) + \mathrm{Var}(\overrightarrow{E}(\hat{\beta}^*)).\end{aligned}$$

This variance can directly be estimated by

$$\overrightarrow{\mathrm{Var}}(\overrightarrow{E}(\hat{\beta} - \hat{\beta}^*) = \frac{1}{R(R-1)} \sum_{r=1}^{R} \{(\hat{\beta}^{(r)} - \hat{\beta}^{*(r)})^2 - R[\overrightarrow{E}(\hat{\beta} - \hat{\beta}^*)]^2\}.$$

As an illustration we reconsider the NLS estimator of β_2 and the unfeasible efficient LS estimator of program mcs34.prg. We know that the OLS estimator is unbiased. The NLS estimator shows a small negative bias which is not significant at a very small significance level. We find that over the simulation replications the two estimators have a

correlation coefficient 0.76. Taking the difference of their expectation estimates yields $0.998698 - 0.999710 = -0.001013$ and this has standard error 0.00034, so we do establish that the bias of NLS in this model, although very tiny, is negative indeed. If we had not applied these two estimators to exactly the same dependent variables y, a much larger number of replications would have been required to find this result (of admittedly very little practical impact). Note, however, that this example in fact establishes an application of a control variate. By the comparison with the unfeasible but unbiased OLS estimator we estimate the bias (std.err.) not to be $-0.001302(0.00052)$, but $-0.001013(0.00034)$, which yields an increase of the MCS efficiency.

In program mcs41.prg an asymptotic test (RESET) and a modified version have been examined under the null and under an alternative hypothesis and rejection probabilities have been estimated. This is done both for $m = 1$ and $m = 2$ and all six probabilities have been estimated using the same R realizations of the disturbance vector. The variables REJR10 and REJR10m have sample correlation 0.85, hence exploiting this it could have been deduced from much fewer replications that the modification reduces the rejection probability. The correlation between REJR20 and REJR20m is 0.78. By generating in EViews the variable REJR20m-REJR20 and calculating its standard error we can establish that the decrease in rejection probability of 0.0284 has a standard error of 0.0017, so the reduction is strongly significant. The correlation between REJR20m and REJR21m is only 0.21 which makes that the standard error of their difference is still 0.0044.

Similar use can be made of the correlation between the MCS estimates in Table 4.3. Its evidence is not fully convincing that at a sample size of $n = 30$ it is useful to modify the LM serial correlation tests to F-form in order to bring its actual significance level closer to the nominal one. However, examining the significance of the difference in rejection probability of LM2 and LMF2 (which is 0.0141) we find that it has standard error 0.0012, and for LM3 and LMF3 this is 0.0101(0.0010). These differences and their strong significance may seem piffling, but they are not. If these rejection probabilities had been estimated independently, then despite their unbiasedness and consistency for large R, they would not have depicted so clearly their systematic differences.

It is the much higher significance of any differences between MCS estimates of moments and probabilities which benefit tremendously when we calculate all the techniques examined for exactly the same draw of the dependent variable in each replication.

The same holds for the quantity which most probably is the MCS result that is most often used for comparisons and ratings of alternative inference techniques, namely the (R)MSE. In Section 2.3 we already disclosed how to assess the standard error of a MCS (R)MSE estimate. Due to its complexity this is seldom done. Even more complex it would be to assess the standard error of a difference in (R)MSE estimates, although such differences are exactly to what most MCS studies boil down to. Undoubtedly these standard errors will be smaller most of the time when the simulation program is designed such that all examined techniques are conditioned on the same R replications of the dependent variable.

5.5 A Comprehensive MCS Methodology

In the initial sections the focus was on MCS methods to analyze particular properties of single estimators or test procedures for just one DGP. In various of the illustrations we already saw that the major interest is often actually on establishing the differences between various alternative competing inference methods over particular families of DGP's or model classes. Then the aim is to rate the qualities of these inference techniques on the basis of the MCS study and next to provide general advice to practitioners. The questions to be answered on the basis of the MCS findings are then: What estimator has smallest (median) bias and smallest MSE (IQR)? Or, what test procedure has best control over actual significance level and has best (size-corrected) power for particular model types? Or, related to that: What procedure for the construction of confidence sets has best control over coverage probability and has least "size-corrected volume"?

Organizing an open championship between alternative inference techniques ("a horse race") requires to set contest rules.[2] Such a

[2] When reading the report on a MCS study one should always be very critical, because in practice, the researcher who did design the study is often not only the organizer of the

tournament should be played within bounds set jointly by particular practical limitations and demands set by impartial scientific ambitions. One should find a way to generate a maximum of useful information, within the limits and requirements set by the available resources, technical tools and appropriate scientific standards. We shall first give some obvious examples of practical limitations and next indicate some rules for a conscientious validation methodology.

The practical limitations are dependent on the available limited computer power, but also on the available limited brain power. A researcher has to allocate her/his time in an optimal way at the one hand to efforts finding satisfying analytical solutions to problems and at the other hand to efforts writing useful and correct MCS programs. These are either very naive and therefore inefficient or they are more efficient because they incorporate elements of analytical findings as in the case of invariance properties and control variates. Consequences of the unavoidable practical limitations are that the class of models examined must have a parameter space with just few dimensions. And in this parameter space we can examine just a limited number of DGPs, that is, of parameter value grid-points. Hence, each DGP has a limited number of ("typical") explanatory (deterministic or random) variables, it can be generated just for a sample size of, say, $n = 10$, 20, 50, 100, 1,000. In dynamic models we can just focus on particular forms of initial conditions. Not just for the parameters of interest, but also for the nuisance parameters, we can only examine a limited grid of values. Tests and confidence sets can be examined just for particular nominal significance levels, say $\alpha = 0.01$, 0.05. And the number of replications and therefore the accuracy of the MCS results cannot escape limitations, but should nevertheless meet particular requirements regarding precision. These are all quantitative restrictions. The methodological requirements are more of a qualitative nature. Here follows our attempt to set some qualitative rules or requirements (or aspirations at least) for the design of a MCS horse race.

---

tournement, but (s)he is also the sponsor and manager of one of the contestants and at the same time acts as the referee.

5.5.1 Ten Commandments of MCS

Methodological rules for an adequate and impartial design of simulation contests that aim to rank various alternative inference techniques are:

1. *Reparametrization of the DGP parameter space into an orthogonal autonomous design parameter space in order to enhance the MCS efficiency and interpretability;*
2. *Explicit exploitation of any invariance, symmetry and proportionality properties;*
3. *Exploration of the non-invariance properties, where possible analytically and for the remaining part experimentally;*
4. *No imposition of dimensional restrictions on the relevant nuisance parameter space;*
5. *Well-argued choice of the examined design parameter values, preferably supported by empirical evidence;*
6. *Each contending technique should play both at home and away, so on each of the other contestants home ground;*
7. *Inclusion of (non-)operational full information techniques in the contest;*
8. *All comparisons of properties of inference techniques should in principle involve conditioning on endogenous variables and arguments should be given where and why conditioning on exogenous variables did (not) occur;*
9. *Test power or confidence set coverage probabilities should be assessed and presented only in conjunction with actual operational and successful size control achievements;*
10. *Reproducibility enabling documentation, including all information required for appropriate interpretation and accuracy assessment.*

Although some of these rules will immediately be obvious, since they refer to opinions developed already in the earlier sections, others may need some further clarification and illustration, which will be given below. Most of these rules are heavily interrelated. Rules 1 through 6 all refer in a way to the set of parameter designs to be covered. Both 6 and 7 address (but from opposite directions) which "horses" (techniques)

should race against each other on which separate "racetracks" (parameter designs). Rules 8 and 9 are about the type of results that should be used in ranking the qualities of alternative techniques, and the 10th commandment entails obvious research quality standards.

Note that in actual practice the quantitative limitations can never be avoided, but within the chosen quantitative restrictions the qualitative requirements can (and thus should) be largely fulfilled!

How useful it is to transform the parameters into autonomous econometric design parameters (rule 1) has already been illustrated in program mcs51.prg, where we made instrument strength and simultaneity severity the major design parameters, reduced at the same time the dimensionality of the parameter space substantially, and found symmetry and proportionality properties (rule 2). The relevance of rules 7 and 8 was already illustrated in program mcs43 and rule 8 again in mcs51. Rule 9 was already substantiated in Section 4.

We addressed the relevance of most of these rules already in Kiviet (2007) in an illustration regarding rating and validating alternative consistent method of moments estimators for simple dynamic panel data models. We will paraphrase that study here.

In panel data the observed variables have more than one dimension, usually two with one related to cross-section units and the other to time period, indicated by i and t, respectively. When the panel data are balanced the same T time periods (years, quarters, weeks, etc.) are covered for all N cross section units (individuals, families, countries, etc.). One of the great attractions of panel data analysis is that when a relationship involves a relevant explanatory variable that has to be omitted because it has not been observed (unobserved heterogeneity) and when this heterogeneity is either time-invariant (IQ, managerial skills, i.e., individual-specific effects) or individual-invariant (global shocks to the economy, i.e., time-specific effects) then the model can be transformed such that the unobserved heterogeneity and its coefficient are removed and either large-N or large-T (or both) consistent estimators can still be found for all or most remaining parameters.

We consider a very specific simple case of dynamic panel data model, viz. the zero-mean fully-stationary normal panel AR(1) model with random unobserved individual-specific time-invariant effects, which for

a balanced panel can be denoted as

$$\left.\begin{array}{l} y_{it} = \gamma y_{i,t-1} + \eta_i + \varepsilon_{it} \\ y_{i0} = \dfrac{1}{1-\gamma}\eta_i + \dfrac{1}{\sqrt{1-\gamma^2}}\varepsilon_{i0} \\ \eta_i \sim NIID(0,\sigma_\eta^2) \\ \varepsilon_i = (\varepsilon_{i0},\varepsilon_{i1},\ldots,\varepsilon_{iT})' \sim NIID(0,\sigma_\varepsilon^2 I_{T+1}) \end{array}\right\} \; i{=}1,\ldots,N;\;\; t{=}1,\ldots,T. \tag{5.6}$$

The parameter σ_ε serves scaling purposes only, so we have in fact just two parameters, viz. γ and $\sigma_\eta/\sigma_\varepsilon$. Due to the chosen stationary start-up values y_{i0}, the two independent and homoskedastic error components η_i and ε_{it} contribute two separate time-invariant and individual-invariant components to

$$\mathrm{Var}(y_{it}) = \frac{\sigma_\eta^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2}. \tag{5.7}$$

If one uses $(\gamma, \sigma_\eta/\sigma_\varepsilon)$ as a base for the design parameter space then keeping $\sigma_\eta/\sigma_\varepsilon$ fixed while varying the dynamics γ will lead to series y_{it} with a different balance regarding the two error components. This seems undesirable, because the dynamics of the relationship is one thing, but to what degree the variance of y_{it} is determined by the two error components is another. An alternative base for the parameter space $\gamma, \sigma_\eta/\sigma_\varepsilon)$ is given either by (γ, ψ), or by (γ, μ), where

$$\psi \equiv \sqrt{\frac{1+\gamma}{1-\gamma}}\frac{\sigma_\eta}{\sigma_\varepsilon},\;\; \mu \equiv \frac{1}{1-\gamma}\frac{\sigma_\eta}{\sigma_\varepsilon}. \tag{5.8}$$

The (γ, ψ)-base allows to change γ (the dynamics) while keeping the ratio of the two variances of the accumulated error components in y_{it} fixed (and vice versa); the (γ, μ)-base achieves this with respect to the ratio of the accumulated individual effects variance and the idiosyncratic disturbance variance. So, both ψ and μ seem closer to autonomous characteristics of the examined population than $\sigma_\eta/\sigma_\varepsilon$. Therefore, it seems that designing a grid for (γ, ψ) or for (γ, μ) is better than for $(\gamma, \sigma_\eta/\sigma_\varepsilon)$, both regarding the interpretation of the results and regarding the efficiency (fewer grid points generate more information). A substantial number of MCS studies in this context did not only

choose $(\gamma, \sigma_\eta/\sigma_\varepsilon)$ as a basis, but also did not vary the parameter $\sigma_\eta/\sigma_\varepsilon$, just giving it value unity. This would be right, if invariance with respect to $\sigma_\eta/\sigma_\varepsilon$ had been established, but this not being the case, keeping it fixed is sinning against rule 4. This may ruin a simulation study. Bun and Kiviet (2006) derived for one of the GMM panel data estimators frequently examined, that

$$E(\hat{\gamma} - \gamma) = -\frac{1}{N} |g(\gamma, \sigma_\eta/\sigma_\varepsilon)| \left(1 - \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}\right) + O\left(\frac{1}{N}\frac{1}{T}\right).$$

Hence, choosing $\sigma_\eta/\sigma_\varepsilon = 1$ in the simulation will give the impression that its bias is much smaller than it actually is when $\sigma_\eta/\sigma_\varepsilon \neq 1$.

Rule 5 urges that when varying the values of parameters like ψ or μ the chosen grid of values should preferably be based on empirical evidence regarding their most likely relevant values. However, in applied papers one does not very often present estimates from which empirically relevant values of $\sigma_\eta/\sigma_\varepsilon$ can be assessed.

Rule 6 refers to the following. If one compares two or more methods in MCS whereas one of them is more robust than the others, it would be unfair to restrict the set of DGP's considered in the various designs to those where the nonrobust methods excel. So when method A is consistent when some regressors are weakly exogenous and method B is inconsistent when regressors are weakly exogenous but only consistent when they are strongly exogenous, whereas in this context the occurrence of weakly exogenous regressors is most likely in practice, then the contest should highlight this limitation of method B and not only its possible superiority under strong exogeneity.

By following rule 7 one will learn what the price is of not having particular information available. Further, it may enable in an informal way to use control variates (as we just illustrated for program mcs43) to estimate this price very precisely.

In the context of the present simple dynamic panel data model rule 8 could imply that one provides arguments why the analysis is made conditional on the assumed double stationarity (with respect to both error components) of the initial condition. Some estimators are robust with respect to this but others exploit it and gain efficiency, whereas they may collapse if this condition were not fulfilled. Another issue here

is again related to conditioning. Should the N initial values be drawn only once, and then be kept constant over all the replications, or should new initial values be drawn every replication? Any study should at least mention what has been done, and why. If the panel should mimic a small set of countries, then it seems reasonable to keep the initial values fixed. But when it represents a microeconometric panel which is supposed to be drawn from a population with many more individuals than N then generating new drawings every replication seems the best option.

Regarding rule 9 we would like to make the following remark. Imagine that for the various possible estimators of the simple panel AR(1) models we would compare the size and power of Wald-type tests on γ. Imagine that over a wide range of γ values, say $0 \leq \gamma \leq 0.9$, a particular technique clearly outperforms all others regarding size-corrected power. If no feasible method is available to do a size correction in practice the outcome of this contest is useless.

Rule 10 is self-evident. It would involve that all variables on which the MCS study has been conditioned should be presented together with the simulation results. In very particular cases reproducibility would (unintentionally) imply that even the actual random numbers used should be provided. We allude to a fault very often made, even in publications that appeared in outlets of the highest reputation (including some by myself): using a MCS sample average as an estimator of a moment which does not exist. In that case the LLN does not apply and the resulting sample average remains random, irrespective of the magnitude of R and a useful interpretation is illusory. In this case (for instance when IV uses just as many instruments as there are unknown coefficients[3]) the median should be used. Oddly enough, especially when R is very small,[4] the probability of a severe outlier is minor and the sample average is often very close to the median, whereas when R is large the MCS sample average behaves as a marsh fire and unless

[3] Very many MCS studies on dynamic panel data techniques contain inept results on the Anderson-Hsiao estimators.

[4] That computing in general and simulation in particular was yet in its infancy two decades back can be illustrated by reconsidering the still widely cited paper by Arellano and Bond (1991), who estimated (several non-existent) moments and rejection probabilities from only $R = 100$ replications.

one uses exactly the same random numbers the obtained same sample average cannot be reproduced.

5.6 Supporting Empirical Findings by a Dedicated MCS Study

It should be clear that certain compromises regarding the aimed at aspirations cannot be avoided when designing a classic MCS study. Its synthetic elements simply cannot be chosen such that adequate information will be generated that is directly applicable to a great many relevant practical situations. A MCS study can just provide a limited view on all the details of the landscape that one is trying to discover. In that respect MCS may be compared with the white stick with red bands, by which a blind person tries to prevent disasters when finding his or her way. Although many details remain unnoticed, when used cautious and deliberate, this stick allows to cope remarkably well amidst heavy traffic and along canals or train platforms with no railings. However, the length of such a stick is limited. So, it just enables to disclose rather precisely the contours of limited local areas. Much detail will remain unnoticed when it is used for more global purposes. Therefore, the major aim should of course be that the set of DGP's examined in a stylized MCS represents the typical elements of actual DGP's for empirical data posses rather well.

Instead of aiming at giving final answers to general questions regarding which technique seems better than others for particular classes of models, one could also choose for the other extreme, and limit the scope of the MCS study by focusing just on examining the relative qualities of various techniques in the very specific special case one has at hand. Then one builds the Monte Carlo design such that it mimics as closely as possible the conjectured DGP of the empirical data under study. This implies using the actual observations on the exogenous regressors and the actual initial conditions. Varying the sample size is of no concern now; one simply sticks to the size of the actual empirical sample. In fact, most of the illustrative programs discussed so far, could easily be adapted in this vein, by simply substituting for the regressors the actually observed explanatories of the relationship under study.

However, choosing numerical values for the parameter values is more problematic.

Of course, one should choose the parameter values in an area covering their suspected true values. It seems obvious that for this choice, and also for the choice of the distribution of the disturbances, one should find inspiration from the empirical estimates. Here one faces the dilemma that one performs a Monte Carlo in the first place because there are doubts regarding the qualities of the estimation technique to be used. Therefore, one should be hesitant to base the MCS design on its findings. However, doing so can be formalized and rationalized, provided these estimators are consistent, and then one enters the world of bootstrapping. This is the topic of the next and final section.

Exercises

1. Consider the standard normal linear regression model with K fixed regressors. Prove that the OLS estimators $\hat{\beta}_j^{(2l-1)}$ and $\hat{\beta}_j^{(2l)}$ $(j = 1, \ldots, K)$ will be negatively correlated when the latter is obtained with disturbance vector $u^{(2l)} = -u^{(2l-1)}$, for $l = 1, \ldots, R/2$. Derive the effects for $E[\overrightarrow{E}(\hat{\beta}_j)]$ and $\mathrm{Var}[\overrightarrow{E}(\hat{\beta}_j)]$, assuming that the $u^{(2l-1)}$ drawings are independent for $l = 1, \ldots, R/2$. Adapt program mcs22.prg to establish the effect on the improved accuracy of $\overrightarrow{E}(\hat{\beta}_j)$ experimentally. By what proportion could one reduce R now to achieve the same accuracy as by the original program?
2. Adapt program msc51.prg such that the magnitude of the effect of conditioning on either z_i or on both z_i and w_i is examined when the instrument is much stronger.
3. Also examine the effects of (not) standardizing both z and w in mcs51.prg for a strong/weak instrument and mild/severe simultaneity.
4. Examine the claim that the density of the IV estimation error is the mirror image when changing the sign of ρ_{xu} and is unaffected by changing the sign of ρ_{zx}. Does this still hold when some of the series u_i, z_i, and w_i have a nonsymmetric distribution?

5. Incorporate in program mcs51.prg a second and a third instrumental variable in the reduced form so that you can control the strength of all instruments separately. Next, use the program to compare the RMSE of both IV and OLS (not conditioning on the exogenous variables) and find values for the weakness of the instruments where OLS beats IV. Check the effect of (not) conditioning on the dependent variable.
6. Adapt program mcs42.prg such that you focus just on finding any serious differences between actual and nominal significance level of the tests LMt1R, LMt1, LM2, and LM2F. Try (in the light of the 10 commandments) to come to more general conclusions that hold for the family of DGP's where the single exogenous regressor is: (a) a stationary AR(1) process with serial correlation coefficient 0.99; (b) a random walk with zero drift; (c) white noise. Provide arguments for your choice of MCS design and the chosen grid of parameter values (what n; what α; what y_0; do you keep the x_t series constant or not?; how do you control the signal to noise ratio of the regression?).
7. Write a program to examine the effects on bias and RMSE of estimating a regression model with first-order serial correlation either by OLS (neglecting serial correlation), GLS (unfeasible), FGLS (by estimating the serial correlation in one way or another) and OLS (after including the first lag of the regressors and the regressand in the model). What is the price of not knowing the serial correlation coefficient? What are the effects on the coverage probability of a confidence set on a slope coefficient constructed in the usual way?

6

Monte Carlo Tests and Bootstrap Inference

In Section 4 we saw that when the null distribution of a test is pivotal in finite samples, that also when this null distribution is nonstandard its quantiles can be assessed with great precision by a MCS study involving all the determining deterministic typical characteristics of the DGP, so that in principle an exact test can be obtained. Here we will show that difficulties around finding the quantile with great precision can be avoided, because a fully exact test and its exact p-values can also be obtained even when in this MCS the number of replications is chosen very small. This already well-established randomization technique is called Monte Carlo testing. Nowadays it is seen as a special fully parametric variant of bootstrapping.

In this final section we will demonstrate more generally how, by incorporating into the design of a MCS more and more typical aspects of a particular empirical inference problem, a classic MCS study may gradually mutate into the generation of bootstrap inference. Below we will indicate more explicitly the various stages from which a MCS study of a fully synthetic DGP may evolve from a dedicated MCS study built around a particular empirical problem, via the generation of pivotal Monte Carlo or parametric bootstrap inference, toward a fully empirical semi-parametric bootstrapping exercise.

6.1 Pivotal Monte Carlo Tests

Let the null-distribution of a continuously distributed test statistic q_n be such that we can simulate it, hence it is pivotal. Using the notation of Section 4, it then has a density $f^{0*}_{q_n}(q; X, y_0, \theta_{10})$ not determined by nuisance parameters. Let $\hat{q}_n$ denote the observed test statistic from the empirical sample. Again we assume that the test has been constructed such that H_0 should be rejected when $\hat{q}_n$ is large. IID drawings $q_n^{(j)}$, $j = 1, \ldots, B$, where B is a finite integer number, can be generated by simulation using the assumed type of distribution of the disturbances, and the empirical observations on X, y_0 and the values θ_{10} of the tested hypothesis. Of course, the actual value $\hat{q}_n$ is independent of these generated $q_n^{(j)}$ drawings, which we can sort in increasing order, yielding $q_n^{*(1)} \leq \cdots \leq q_n^{*(B)}$. These B values define $B + 1$ disjunct intervals: $(-\infty, q_n^{*(1)}]$, $(q_n^{*(1)}, q_n^{*(2)}]$, $\ldots, (q_n^{*(B)}, \infty)$. Assuming that the null hypothesis is true, so $\hat{q}_n$ has been drawn from the same distribution as the $q_n^{(j)}$, we have $\Pr(\hat{q}_n \leq q_n^{*(j)}) = j/(B + 1)$. Using $q_n^{*(J)}$ as critical value for test statistic $\hat{q}_n$, where $J \in \{1, \ldots, B\}$, yields rejection probability

$$\Pr(\hat{q}_n > q_n^{*(J)}) = 1 - \frac{J}{B+1} = \frac{B+1-J}{B+1}. \tag{6.1}$$

This test procedure is exact with significance level α if $(B + 1 - J)/(B + 1) \leq \alpha$, so for J we should choose the smallest integer that satisfies $J \geq (1 - \alpha)(B + 1)$.

Ideally, α and B are such that $\alpha(B + 1)$ is integer and then the test is similar with $\Pr(\hat{q}_n > q_n^{*(J)}) = \alpha$. Thus, for $\alpha = 0.05$, an appropriate choice for the Monte Carlo sample size would be $B = 999$, giving $J = (B + 1)(1 - \alpha) = 1000 \times 0.95 = 950$. However, $B = 199$ $(J = 190)$ or 99 $(J = 95)$ or even $B = J = 19$ achieve for this test similarity at $\alpha = 0.05$ too, on top of exactness. In this way we have randomized the test procedure, and doing it in this way is known as "Monte Carlo testing." Such procedures date back to the 1950/1960s; see, for instance, Dufour and Khalaf (2001) for further references. These days they are often categorized as a special kind of exact parametric bootstrap test procedure.

The Monte Carlo test procedure based on test statistic $\hat{q}_n$ and an arbitrary number of replications B has p-value

$$p_B(\hat{q}_n) = \frac{1}{B}\sum_{j=1}^{B}\mathbb{I}(q_n^{(j)} > \hat{q}_n) = \frac{1}{B}\sum_{j=1}^{B}\mathbb{I}(q_n^{*(j)} > \hat{q}_n), \tag{6.2}$$

where $\mathbb{I}(\cdot)$ is again the indicator function. So, for calculating p-values there is no special advantage in choosing B equal to values like 19, 99 or 999. Because the p-value of a test statistic contains more information than just its significance (yes or no) at a particular level, it is better to focus on calculating p-values and then choosing values for B which are powers of 10 (and at least 100) in order to avoid unnecessary truncation.

Because the above procedure can be generalized in a particular way for test statistics with a nonpivotal null distribution, we best refer to it as the PMC (pivotal Monte Carlo) test procedure. It is exact (for both n and B finite!), but has the curious property that when two researchers use it on the same model and data set, they may obtain conflicting inference (different p-values, thus one of them may reject at a particular level and the other not), due to the randomness of the Monte Carlo. This would be avoided for $B \to \infty$, because for any J the value of $q_n^{*(J)}$ will converge to the true $J/(B+1)$-quantile of the null distribution and thus for $J = (1-\alpha)(B+1)$ to the $(1-\alpha)$-quantile. Also the power of the procedure will benefit from choosing a large B, but it has been found that in general a value of B exceeding 1,000 does not lead to much power improvement.

6.2 Implementations of PMC Tests

Here we provide a few examples of PMC tests. Note that a null distribution can only be pivotal in finite samples if one is willing to make an assumption on the actual type of the distribution of the disturbances, up to a scaling factor (σ). In all illustrations below, we choose normality for this. However, it would be straight-forward to convert the programmes and adopt any other type of distribution, as long as this would not involve an extra unknown parameter. So, we could go for disturbances with fatter tails than the normal and take the Student distribution for that, but then we also would have to make an assumption

on its number of degrees of freedom (lower for fatter tails). Also skew disturbances would be possible, but not without pre-setting the actual skewness coefficient and those of the higher-order moments. The second subsection below makes clear though that one only has to specify the distribution fully under the null hypothesis, and therefore a test for normality of the disturbances can also be cast into the framework of a PMC test.

6.2.1 Exact Similar Testing for Serial Correlation in Classic Regression

In the linear normal regression model with fixed regressors

$$y = X\beta + u, \tag{6.3}$$

where X is $n \times k$, we will compare the situation $u \sim N(0, \sigma^2 I_n)$ with $u \sim N(0, \sigma^2 \Omega)$, where $\Omega = (\omega_{ij})$ with $\omega_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, n$ and $\rho \neq 0$. One can test $H_0 : \rho = 0$ against $H_1 : \rho > 0$ by the Durbin–Watson test using bounds to the critical values. These have been tabulated for different values of n and $k' = k - 1$ (assuming there is an intercept in the regression) and for specific significance levels α. These bounds are used, not because the test statistic is not a pivot under the null hypothesis (as we will show, the null distribution is invariant with respect to both β and σ), but because the null distribution is not invariant with respect to X. Hence, exact critical values cannot be tabulated, because that would require a separate table for each possible X matrix (in fact, it is just the space spanned by X that determines the null distribution). Using the lower bound critical value makes the test exact but conservative and nonsimilar because the actual type I error probability will vary with X. However, it is pretty straightforward to cast this test into the framework of Monte Carlo testing and thus avoid the inconvenience (and power loss!) of the bounding critical values, and obtain an exact and even a similar test procedure. This is done as follows.

The Durbin–Watson statistic is defined as

$$d \equiv \frac{\sum_{i=2}^{n} (\hat{u}_i - \hat{u}_{i-1})^2}{\sum_{i=1}^{n} \hat{u}_i^2} = \frac{\hat{u}' A \hat{u}}{\hat{u}' \hat{u}}, \tag{6.4}$$

where $\hat{u} = y - X\hat{\beta} = [I - X(X'X)^{-1}X']y = M_X u$ and A is an $n \times n$ matrix where $A = (a_{ij})$ with $a_{11} = a_{nn} = 1$, $a_{ii} = 2$ for $i = 2, \ldots, n-1$ and $a_{i,i+1} = a_{i+1,i} = -1$ for $i = 1, \ldots, n-1$. Hence, we have

$$d = \frac{u' M_X A M_X u}{u' M_X u} = \frac{\varepsilon' M_X A M_X \varepsilon}{\varepsilon' M_X \varepsilon}, \tag{6.5}$$

with $u = \sigma\varepsilon$, and $\varepsilon \sim N(0, I_n)$ under H_0. Thus, the null distribution of d does not involve any unknown parameters, although it does depend on the space spanned by the columns of X through M_X and on the assumed normality of the disturbances. It is straightforward to simulate drawings from the null distribution of d. We do so in the illustrating program mcs61.prg which compares in a simple empirical example the use of and difference between the bounds test version of the Durbin–Watson test and its PMC test version. The file ICECREAM.wf1 contains data taken from Hildreth and Lu's 1960 paper on "Demand relations with autocorrelated disturbances" (Technical Bulletin 276, Michigan State University). The data used in this study are time series data with four-weekly observations from 18 March 1951 to 11 July 1953 on the following variables:

```
cons:     consumption of ice cream per head (in pints)
income:   average family income per week (in US Dollars)
price:    price of ice cream (per pint)
temp:     average temperature (in Fahrenheit)
```

Consider the following EViews program:

```
'mcs61.prg:   Durbin-Watson statistic used with bounds and as Monte Carlo test
load f:\MCS\icecream.wf1
genr priceincome=price*income
equation eq1.ls cons c income price temp priceincome
!dwemp=eq1.@dw
!B=10000
vector (!B,1) simdw
rndseed 9876543210
for !rep=1 to !B
   genr eps=nrnd
   equation eq2.ls eps c income price temp priceincome
   simdw(!rep,1)=eq2.@dw
next
simdw.write f:\MCS\mcs61sim.txt
```

```
workfile f:\MCS\mcs61sim.wf1 u 1 !B
read f:\MCS\mcs61sim.txt dwsim
genr smaller=dwsim<!dwemp
```

We find for eq1:
Dependent Variable: CONS

	Coefficient	Std. Error	t-Statistic	p-value
intercept	−6.910	3.48	−1.98	0.06
income	0.087	0.041	2.13	0.04
price	25.26	12.87	1.96	0.06
temp	0.0031	0.0004	6.85	0.00
priceincome	−0.308	0.151	−2.04	0.05
R-squared	0.759			
Durbin–Watson stat	1.24			
Included observations	30			

Because at 5% $d^L = 1.14$ and $d^U = 1.74$ the bounds test is inconclusive. However, Figure 6.1 shows that the Monte Carlo test version yields for the variable "smaller" a mean of 0.0019. Hence, the p-value is so small that we reject the hypothesis of no serial correlation. For the null distribution of the *dw* statistic for this particular X matrix under normality we find the empirical PDF as represented by the histogram of variable *dw*sim.

Note how the program makes use of the fact that for generating the null distribution of the test statistic we do not need values for β nor for σ. Therefore, we simply run B regressions of $\varepsilon \sim N(0, I)$ on the regressors X, in order to obtain R drawings of (6.5).

6.2.2 A PMC Test for Normality

The Jarque–Bera test for normality of the disturbances in a regression is an asymptotic test, in the sense that its null distribution is χ^2

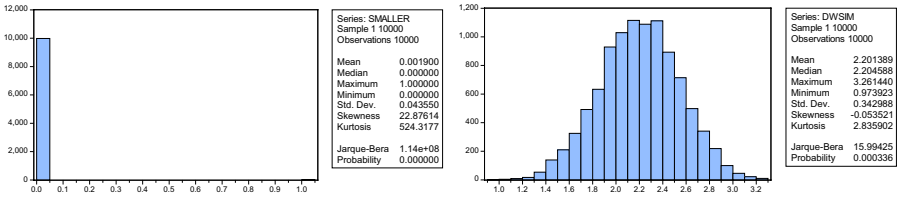

Fig. 6.1 Results for program mcs61.prg.

with two degrees of freedom only asymptotically. However, in a classic regression model where all regressors are exogenous it is easy to use it as a PMC test, because its null distribution is pivotal also in finite samples. The JB test statistic is a nonlinear function of the scaled zero-mean regression residuals $\hat{u}_i/s$, $i = 1, \ldots, n$, where $\hat{u} = y - X\hat{\beta} = M_X u$, with $M_X = I_n - X(X'X)^{-1}X'$, and $s^2 = \hat{u}'\hat{u}/(n-k)$, where k is the number of columns of the full column rank regressor matrix X and $u = (u_1, \ldots, u_n)'$ is the vector of disturbances with $u_i \sim IID(0, \sigma^2)$. Under the null hypothesis of the Jarque–Bera test we have $u_i = \sigma\varepsilon_i \sim NIID(0, \sigma^2)$. The subroutine Jarque–Bera in program mcs62.prg calculates

$$JB = \frac{n-k}{6}\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{u}_i}{s}\right)^3\right]^2 + \frac{1}{4}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{u}_i}{s}\right)^4 - 3\right]^2\right\}. \qquad (6.6)$$

Note that

$$\frac{1}{s}\hat{u} = \left(\frac{n-k}{u'M_X u}\right)^{1/2} M_X u = \left(\frac{n-k}{\varepsilon' M_X \varepsilon}\right)^{1/2} M_X \varepsilon,$$

so under the null hypothesis JB is pivotal indeed. Note (and explain!) that in the program below, like in the foregoing one, a regression of a series of generated standard normal variables is run on the regressors X, in order to obtain drawings from the genuine null distribution of the test.

```
'mcs62.prg:  Jarque-Bera test used as a Monte Carlo test
 subroutine JarqueBera(series uhat, scalar n, scalar k, scalar jb)
  genr scaleduhat = uhat/@sqrt(@sum(uhat^2)/(n-k))
  genr scaleduhat3 = scaleduhat^3
  genr scaleduhat4 = scaleduhat^4
  jb = (n-k)*((@sum(scaleduhat3)/n)^2 + 0.25*((@sum(scaleduhat4)/n)-3)^2)/6
endsub
load f:\MCS\icecream.wf1
genr priceincome=price*income
equation eq1.ls cons c income price temp priceincome
eq1.makeresids empresid
!n = eq1.@regobs
!k = eq1.@ncoef
scalar jb
call JarqueBera(empresid, !n, !k, jb)
!jbemp=jb
```

```
!B=10000
vector (!B) simjb
rndseed 9876243210
for !rep=1 to !B
   genr eps=nrnd
   equation eq2.ls eps c income price temp priceincome
   eq2.makeresids simresid
   call JarqueBera(simresid, !n, !k, jb)
   simjb(!rep)=jb
next
simjb.write f:\MCS\mcs62sim.txt
workfile f:\MCS\mcs62sim.wf1 u 1 !B
read f:\MCS\mcs62sim.txt jbsim
genr larger=jbsim>!jbemp
```

The empirical realization of the test statistic obtained by EViews is 1.47, and its associated asymptotic p-value is 0.48. The p-value of its PMC implementation can be found in Figure 6.2; it is 0.77. However, for the JB value calculated with degrees of freedom correction according to (6.6) we find 1.185 (the value of !jbemp in the program) for these empirical data (see also Exercise 5). This value has been used in establishing the series "larger." Normality is not rejected, neither by the asymptotic test, nor by the exact test, but note that the power of the test cannot be very large in a sample of such a small size.

6.2.3 Exact Tests in ARX(1) Models

Because the null distributions of the Durbin-Watson and of the Jarque–Bera test statistics are already pivotal by nature, casting them into the framework of PMC tests to obtain exact inference is relatively easy. Usually, test statistics have nonpivotal null distributions, and then the test statistic should be adapted to enable exact inference. This seems

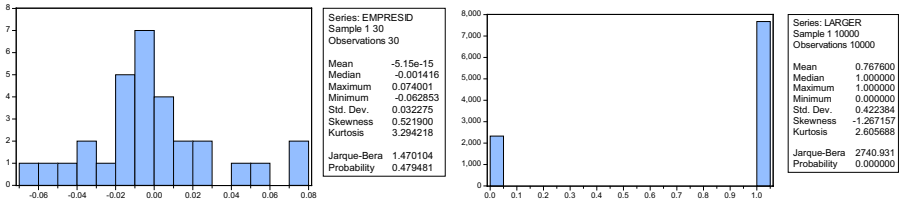

Fig. 6.2 Results for program mcs62.prg.

impossible in most cases, but sometimes it can be done, as we illustrate below for generalized forms of particular classic unit root tests.

The usual test for $H_0 : \gamma = \gamma_0$ in the ARX(1) model $y = \gamma y_{-1} + X\beta + \sigma\varepsilon$, where $\varepsilon \sim (0, I_n)$, is asymptotic, also when the disturbances are normal. The t-test statistic is pivotal under the null only asymptotically. However, an alternative test statistic with pivotal null distribution in finite samples can be found. We first examine why the standard test is nonpivotal. The OLS estimators for γ and β can be expressed as

$$\hat{\gamma} = (y'_{-1}M_X y_{-1})^{-1} y'_{-1}M_X y = \gamma + \sigma \frac{y'_{-1}M_X\varepsilon}{y'_{-1}M_X y_{-1}}, \tag{6.7}$$

$$\hat{\beta} = (X'X)^{-1}X'(y - \hat{\gamma}y_{-1}). \tag{6.8}$$

For the least-squares residuals we find $\hat{u} = y - \hat{\gamma}y_{-1} - X\hat{\beta} = y - \hat{\gamma}y_{-1} - X(X'X)^{-1}X'(y - \hat{\gamma}y_{-1}) = M_X(y - \hat{\gamma}y_{-1})$. Now substituting for y the right-hand side of the ARX(1) model, we obtain $\hat{u} = M_X(\gamma y_{-1} + X\beta + \sigma\varepsilon - \hat{\gamma}y_{-1}) = \sigma M_X\varepsilon - (\hat{\gamma} - \gamma)M_X y_{-1}$, which yields

$$\hat{u} = \sigma M_X\varepsilon - \sigma \frac{y'_{-1}M_X\varepsilon}{y'_{-1}M_X y_{-1}} M_X y_{-1},$$

from which

$$\hat{u}'\hat{u} = \sigma^2\varepsilon' M_X\varepsilon - \sigma^2 \frac{(y'_{-1}M_X\varepsilon)^2}{y'_{-1}M_X y_{-1}} \tag{6.9}$$

follows. For testing $H_0 : \gamma = \gamma_0$, against either $H_1^L : \gamma < \gamma_0$ or $H_1^R : \gamma > \gamma_0$, the familiar asymptotic test statistic is

$$q_\gamma = \frac{\hat{\gamma} - \gamma_0}{\hat{\sigma}(\hat{\gamma})}, \tag{6.10}$$

where

$$\begin{aligned}\hat{\sigma}(\hat{\gamma}) &= \left[\frac{\hat{u}'\hat{u}}{n}\left(y'_{-1}M_X y_{-1}\right)^{-1}\right]^{1/2} \\ &= \frac{\sigma}{\sqrt{n}}\left[\frac{\varepsilon' M_X\varepsilon}{y'_{-1}M_X y_{-1}} - \left(\frac{y'_{-1}M_X\varepsilon}{y'_{-1}M_X y_{-1}}\right)^2\right]^{1/2}.\end{aligned}$$

Because

$$\hat{\gamma} - \gamma_0 = (\gamma - \gamma_0) + \sigma \frac{y'_{-1} M_X \varepsilon}{y'_{-1} M_X y_{-1}},$$

we have

$$q_\gamma = \sqrt{n} \left[\frac{\varepsilon' M_X \varepsilon}{y'_{-1} M_X y_{-1}} - \left(\frac{y'_{-1} M_X \varepsilon}{y'_{-1} M_X y_{-1}} \right)^2 \right]^{-1/2} \left(\frac{\gamma - \gamma_0}{\sigma} + \frac{y'_{-1} M_X \varepsilon}{y'_{-1} M_X y_{-1}} \right),$$

and under $\gamma = \gamma_0$ this specializes to

$$\begin{aligned} q_{\gamma_0} &= \sqrt{n} \left[\frac{\varepsilon' M_X \varepsilon}{y'_{-1} M_X y_{-1}} - \left(\frac{y'_{-1} M_X \varepsilon}{y'_{-1} M_X y_{-1}} \right)^2 \right]^{-1/2} \left(\frac{y'_{-1} M_X \varepsilon}{y'_{-1} M_X y_{-1}} \right) \\ &= \sqrt{n} \left[\frac{(y'_{-1} M_X y_{-1})(\varepsilon' M_X \varepsilon)}{(y'_{-1} M_X \varepsilon)^2} - 1 \right]^{-1/2}. \end{aligned} \tag{6.11}$$

To examine the nuisance parameters of the null-distribution of q_γ, i.e., of q_{γ_0}, we express $M_X y_{-1}$ in terms of the parameters and conditioning variables. Successive substitution yields

$$\begin{aligned} y_{-1} = \begin{pmatrix} y_0 \\ y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_{n-1} \end{pmatrix} &= \begin{pmatrix} 1 \\ \gamma \\ \gamma^2 \\ \cdot \\ \cdot \\ \gamma^{n-1} \end{pmatrix} y_0 \\ &+ \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & & & & \cdot \\ \gamma & 1 & \cdot & & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot & \cdot \\ \gamma^{n-2} & \gamma^{n-3} & \cdot & \cdot & 1 & 0 \end{pmatrix} \begin{pmatrix} x'_1\beta + \sigma\varepsilon_1 \\ x'_2\beta + \sigma\varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ x'_n\beta + \sigma\varepsilon_n \end{pmatrix} \end{aligned}$$

or, upon defining the vector $\iota(\gamma)$ and the matrix $C(\gamma)$,

$$y_{-1} = \iota(\gamma) y_0 + C(\gamma) X\beta + \sigma C(\gamma)\varepsilon.$$

Hence,

$$M_X y_{-1} = M_X \iota(\gamma) y_0 + M_X C(\gamma) X\beta + \sigma M_X C(\gamma)\varepsilon,$$

and under $\gamma = \gamma_0$

$$M_X y_{-1} = M_X \iota(\gamma_0) y_0 + M_X C(\gamma_0) X\beta + \sigma M_X C(\gamma_0)\varepsilon,$$

Thus, the null-distribution of q_γ, which is determined by (6.11), depends in general on

$$\beta,\ \sigma,\ y_0,\ \gamma_0,\ X \text{ (including } T,\ k) \text{ and the actual distribution of } \varepsilon,$$

i.e., the test is nonpivotal. Under the alternative, γ determines the distribution of q_γ too.

Its nonpivotal nature originates from the complexity under $\gamma = \gamma_0$ of

$$M_X y_{-1} = M_X[\iota(\gamma_0) y_0 + C(\gamma_0) X\beta] + \sigma M_X C(\gamma_0)\varepsilon.$$

From this an easy procedure to make the test statistic independent from nuisance parameters emerges immediately. If we regress y not just on y_{-1} and X, but also on the redundant fixed regressors $\iota(\gamma_0)$ and $C(\gamma_0)X$ then, using the notation

$$X^* = [X\ \iota(\gamma_0)\ C(\gamma_0)X],$$

and running the regression of y on y_{-1} and X^*, which gives estimators $\hat{\gamma}^*$, $\hat{\beta}^*$ and residuals $\hat{u}^*$, we find that the test statistic

$$q^*_\gamma = \frac{\hat{\gamma}^* - \gamma_0}{\left[\frac{\hat{u}^{*\prime}\hat{u}^*}{T}\left(y'_{-1} M_{X^*} y_{-1}\right)^{-1}\right]^{1/2}}$$

has no nuisance parameters under $\gamma = \gamma_0$, since

$$\begin{aligned} M_{X^*} y_{-1} &= M_{X^*}[\iota(\gamma_0) y_0 + C(\gamma_0) X\beta + \sigma C(\gamma_0)\varepsilon] \\ &= \sigma M_{X^*} C(\gamma_0)\varepsilon. \end{aligned}$$

Therefore we find

$$\begin{aligned} q^*_{\gamma_0} &= \sqrt{n}\left[\frac{(y'_{-1} M_{X^*} y_{-1})(\varepsilon' M_{X^*}\varepsilon)}{(y'_{-1} M_{X^*}\varepsilon)^2} - 1\right]^{-1/2} \\ &= \sqrt{n}\left[\frac{\sigma^2[\varepsilon' C(\gamma_0)' M_{X^*} C(\gamma_0)\varepsilon](\varepsilon' M_{X^*}\varepsilon)}{\sigma^2(\varepsilon' C(\gamma_0)' M_{X^*}\varepsilon)^2} - 1\right]^{-1/2} \\ &= \sqrt{n}\left[\frac{[\varepsilon' C(\gamma_0)' M_{X^*} C(\gamma_0)\varepsilon](\varepsilon' M_{X^*}\varepsilon)}{(\varepsilon' C(\gamma_0)' M_{X^*}\varepsilon)^2} - 1\right]^{-1/2}, \end{aligned}$$

which is pivotal when the distribution of ε is known: its distribution only depends on ε and the observables X and γ_0. The distribution, though intractable analytically, can easily be simulated. Note also that when running such a test in finite samples it is not an issue whether γ_0 is inside, on, or outside the unit circle. The latter is an important complicating issue for asymptotic tests.

Specific examples of this test are quite well known as Dickey–Fuller tests. Consider the q_γ test for the special case $\gamma_0 = 1$ (unit root) and: (i) no X matrix, $y_0 = 0$; (ii) no X matrix, y_0 arbitrary; (iii) only an intercept, i.e., $X = \iota(1)$, y_0 arbitrary; (iv) only an intercept and linear trend, i.e., $x_t = (1, t)$, y_0 arbitrary.

In (i):

$$q_1^* = \sqrt{n}\left[\frac{(y'_{-1}y_{-1})(\varepsilon'\varepsilon)}{(y'_{-1}\varepsilon)^2} - 1\right]^{-1/2} = \sqrt{n}\left[\frac{(\varepsilon' C(1)' C(1)\varepsilon)(\varepsilon'\varepsilon)}{(\varepsilon' C(1)'\varepsilon)^2} - 1\right]^{-1/2}$$

is pivotal (but not asymptotically normal!). In all other cases q_1 is not pivotal, but q_1^* is. In (ii), for y_0 arbitrary, the redundant regressors $\iota(1)$, i.e., the constant, has to be included. In (iii), where $X = \iota(1)$, the extended regressor matrix X^* should also include $C(1)\iota(1)$, which is the linear trend. In (iv), where $x_t = (1, t)$, one should take $x_t^* = (1, t, t^2)$ in order to obtain a pivotal test statistic, etc. All these specific pivotal test statistics are NOT normal asymptotically, but follow nonstandard (so-called Dickey–Fuller) distributions. Implemented for the pivotal test statistic q_1^* they are exact for any (stable or unstable) value of γ_0 and any assumed pivotal distribution of ε and do not involve any asymptotic approximations. However, they are still asymptotically valid if the distribution of ε is misspecified (say as normal whereas it is actually nonnormal, or probably as IID whereas they are actually serially correlated though stationary).

6.3 Parametric and Nonparametric Bootstrap Inference

By bootstrapping one tries to improve first-order asymptotic inference as follows. Whereas in classic Monte Carlo simulation one simulates a fully parametrized DGP that represents the relevant characteristics of a (usually stylized) population, by bootstrapping one simulates

from an actual empirical sample (this is called resampling). Its simulation results can be used to mitigate estimator bias, or to obtain improved variance estimators, to obtain better critical values for tests and more accurate p-value estimates, and also for confidence sets with more accurate coverage probability than those obtained by straightforward asymptotics. The observed empirical sample is in fact used as the population from which simulations are obtained. In general, only for large empirical samples ($n \to \infty$) this contains all relevant information on the population. Therefore, apart from particular very special situations, bootstrap inference is not exact or unbiased, but still asymptotic although often its approximation errors can be shown to be of smaller order in n than those of standard first-order asymptotic inference.

The bootstrap was developed in the 1970s. Particular forms of it in the context of testing are similar to pivotal Monte Carlo testing, but many more implementations are possible and new ones are still being developed. Its name refers to a tale in which the hero pulls himself up from a swamp by the straps of his boots.

There is a distinction between parametric and semi-parametric bootstrap inference. In the former a full parametric specification is required of the model specification, whereas the latter usually leaves the actual type of distribution of the observed dependent variables (or underlying disturbances) unspecified. When the test statistic is pivotal, the parametric bootstrap test procedure gets close to pivotal Monte Carlo testing. However, bootstrap testing just requires a test statistic that is asymptotically pivotal under the null hypothesis. Therefore, it is applicable to most tests encountered in econometrics (provided an appropriate resampling scheme can be designed) because most tests have an asymptotic null distribution that is either χ^2 or standard normal. When the test is just asymptotically pivotal, also a bootstrap test is only asymptotically valid; then size problems may be mitigated in finite samples by bootstrapping, but the resulting bootstrap test is not exact.

The distinguishing characteristic between alternative bootstrap procedures is the employed resampling scheme. We will illustrate this by a simple example, for which we again take testing in ARX(1) models, because this makes it easy to indicate in the context just exemplified the

distinctive features and general requirements of: (i) standard asymptotic tests, (ii) exact PMC tests, (iii) parametric bootstrap tests and (iv) semi- and nonparametric bootstrap tests. The former two have been illustrated before, whereas the latter two will be demonstrated in the next subsection and also more formally defined in the final section.

6.3.1 Quality of Bootstrap Inference in ARX(1) Models

The model is specified as

$$y_i = \gamma y_{i-1} + x_i'\beta + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

with in the fully parametric model

$$\varepsilon_i \sim NIID(0,1),$$

or any other explicit distribution function, and in the semi-parametric model the distribution function is left unspecified, i.e.,

$$\varepsilon_i \sim IID(0,1).$$

We consider testing $H_0 : \gamma = \gamma_0$ and have obtained the usual asymptotic (studentized) test statistic, which we denote as $\hat{q}_\gamma$. Now B pseudo samples of size n of this regression model are generated substituting the empirical estimates of the parameters (as obtained under the tested H_0), and drawing pseudo-disturbances.

In the parametric bootstrap these pseudo-disturbances are drawn from the specified normal (or other) distribution, and resampling works as follows. Let $\tilde{\beta}$ and $\tilde{\sigma}$ be the (restricted) OLS estimators of β and σ obtained for regression model $y_i - \gamma_0 y_{i-1} = x_i'\beta + \sigma\varepsilon_i$, and let $\tilde{u}_i = y_i - \gamma_0 y_{i-1} - x_i'\tilde{\beta}$ be the (restricted) residuals. Now we generate for $j = 1, \ldots, B$ the "resampled artificial data"

$$\begin{aligned} y_i^{(j)} &= \gamma_0 y_{i-1}^{(j)} + x_i'\tilde{\beta} + \tilde{\sigma}\varepsilon_i^{(j)}, \quad y_0^{(j)} = y_0, \\ \varepsilon_i^{(j)} &\sim NIID(0,1), \quad i = 1, \ldots, n. \end{aligned}$$

and calculate from each sample the test statistic $q_\gamma^{(j)}$, which establishes the empirical null distribution with which we will confront the empirical value $\hat{q}_\gamma$.

In the semi-parametric bootstrap the resampling scheme involves

$$y_i^{(j)} = \gamma_0 y_{i-1}^{(j)} + x_i'\tilde{\beta} + w\tilde{u}_i^{(j)}, \quad y_0^{(j)} = y_0,$$

with $\tilde{u}_i^{(j)}$ drawn randomly with replacement from $(\tilde{u}_1, \ldots, \tilde{u}_n)$, employing a particular weight w, for instance $w = [n/(n - \tilde{k})]^{1/2}$, where $\tilde{k}$ is the number of coefficients in the restricted model. This weight compensates for residuals being on average smaller than disturbances.

The EViews program mcs63.prg produces empirical inference, both standard asymptotic and parametric bootstrap inference. It uses data from a workfile phillips.wf1, which is extracted from Greene (2012, Example 20.3), where a simplified version of the Phillips curve is analyzed on the basis of quarterly US data ranging from 1950 until 2000. We estimate the model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \beta_4 x_{t-1} + u_t,$$

where y_t is the first difference of the inflation rate and x_t the unemployment rate. We test $H_0 : \beta_2\beta_3 + \beta_4 = 0$. The implication of H_0 would be that the specification is actually static with AR(1) disturbances. The program is:

```
'mcs63.prg  parametric bootstrap test
load f:\MCS\phillips.wf1
smpl 1 204
genr pary=y
smpl 7 204
equation eq1.ls y c x y(-1) x(-1)
eq1.wald c(2)*c(3)+c(4)=0
coef (4) cc
equation eq2.ls y=cc(1)+cc(2)*x+cc(3)*y(-1)-cc(2)*cc(3)*x(-1)
!testemp=2*(eq1.@logl-eq2.@logl)
cc(4)=!testemp
!B=1000
rndseed 9876543210
vector (!B) results
for !res=1 to !B
  genr pary=cc(1)+cc(2)*x+cc(3)*pary(-1)-cc(2)*cc(3)*x(-1)+eq2.@se*nrnd
  equation eq3.ls pary c x pary(-1) x(-1)
  equation eq4.ls pary=c(1)+c(2)*x+c(3)*pary(-1)-c(2)*c(3)*x(-1)
  results(!res)=2*(eq3.@logl-eq4.@logl)
next
```

```
results.write f:\MCS\mcs63.txt
workfile f:\MCS\mcs63.wf1 u 1 !B
read f:\MCS\mcs63.txt testpar
genr largerpar=testpar>!testemp
```

In order to be able to compare the estimates and the likelihood later with those when higher order lags are included, we estimate over the observations 7 through 204. For the usual Wald and the LR (likelihood ratio) statistics for testing this nonlinear restriction we find 5.83 and 6.02 with asymptotic p-values 0.016 and 0.014 respectively, which both strongly reject when used as ordinary asymptotic test procedures. The parametric bootstrap assessment of the empirical null distribution of the LR-test follows from Figure 6.3, which does not seem to differ considerably from $\chi^2(1)$. The mean of the variable largerpar is 0.017, hence the bootstrap p-value is remarkably close to the asymptotic one.

Just a few lines of the program have to be adapted in order to perform the semiparametric alternative to this test, see mcs64.prg:

```
'mcs64.prg  semiparametric bootstrap test
load f:\MCS\phillips.wf1
smpl 1 204
genr sempary=y
smpl 7 204
equation eq1.ls y c x y(-1) x(-1)
eq1.wald c(2)*c(3)+c(4)=0
coef (4) cc
equation eq2.ls y=cc(1)+cc(2)*x+cc(3)*y(-1)-cc(2)*cc(3)*x(-1)
!testemp=2*(eq1.@logl-eq2.@logl)
cc(4)=!testemp
genr resreseq2=resid*(eq2.@regobs/(eq2.@regobs-eq2.@ncoef))^0.5
!B=1000
rndseed 9876543210
vector (!B) results
for !res=1 to !B
```

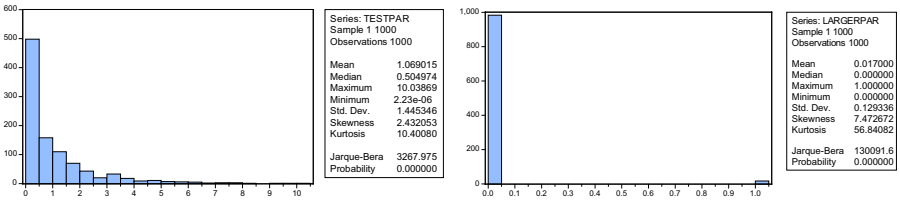

Fig. 6.3 Results from program mcs63.prg.

```
  group g resreseq2
  g.resample
  genr sempary=cc(1)+cc(2)*x+cc(3)*sempary(-1)-cc(2)*cc(3)*x(-1)+resreseq2_B
  equation eq3.ls sempary c x sempary(-1) x(-1)
  equation eq4.ls sempary=c(1)+c(2)*x+c(3)*sempary(-1)-c(2)*c(3)*x(-1)
  results(!res)=2*(eq3.@logl-eq4.@logl)
next
results.write f:\MCS\mcs64.txt
workfile f:\MCS\mcs64.wf1 u 1 !B
read f:\MCS\mcs64.txt testsempar
genr largersempar=testsempar>!testemp
```

The command g.resample resamples the variables in the group g. Through command "group g resresideq2" this group g contains here just the variable resresideq2. The command "g.resample" produces a variable resresideq2_B which contains random drawings with replacement from the elements of resresideq2. Although the residuals from equation 2 seem to be considerably different from normal (see Figure 6.4), the resulting semiparametric bootstrap empirical null distribution of the LR-test does not differ much from its parametric counterpart; now we find a p-value of 0.010.

Of course, the above analysis is built on the assumption that the maintained model is well specified. That is clearly not the case here, because indicating the present unrestricted model as AD(1,1) and testing this against AD(4,4) the LR statistic is $2 \times (453.50 - 424.12) = 58.76$ and this $\chi^2(6)$ statistic has p-value 0.00. Since testing AD(3,3) against AD(4,4) gives $\chi^2(2)$ statistic 3.3 with p-value 0.19 the more parsimonious AD(3,3) specification seems more reasonable. From the above one might get the impression that there is little difference between standard asymptotic and asymptotic bootstrap inference, but that is most possibly due to the fact that the above models are rather

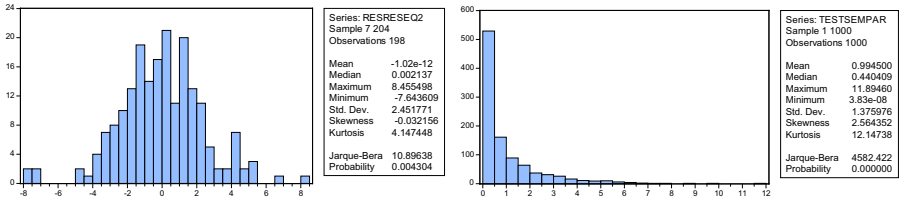

Fig. 6.4 Results from program mcs64.prg.

unusual, because in the AD(1,1) specification the coefficient of the lagged dependent variable is negative.

Therefore, finally we will check by MCS how well the asymptotic technique of bootstrap testing works in finite samples of a dynamic regression model of a nature that occurs more frequently in practice. In program mcs65.prg we therefore again use an artificial DGP, instead of a very specific empirical one. Although we will only analyze one set of values for the coefficients of the DGP, it is of course very easy to examine others as well.

```
'mcs65.prg: MCS on size problems of tests in an ARX(1) model
!n=40
workfile f:\MCS\mcs65.wf1 u 0 !n
!beta1=0
!beta2=0.2
!beta3=1-!beta2
!sigma=0.3
smpl 0 0
genr x=1
genr y=(!beta1+!beta2*x)/(1-!beta3)
genr sempary=y
smpl 1 !n
rndseed 12345
genr x=x(-1)+0.02+0.05*nrnd
!R=1000
!B=999
!w=@sqrt(!n/(!n-2))
matrix (!R,2) simres
vector (!B) bsimres
rndseed 9876543210
for !rep=1 to !R
   genr u=!sigma*nrnd
   genr y=!beta1+!beta2*x+!beta3*y(-1)+u
   equation eq1.ls y c x y(-1)
   simres(!rep,1)=(eq1.@coefs(3)-!beta3)/eq1.@stderrs(3)
   equation eq2.ls y-!beta3*y(-1) c x
   genr resid2=resid
   for !res=1 to !B
     group g resid2
     g.resample
     genr sempary=eq2.@coefs(1)+eq2.@coefs(2)*x+!beta3*sempary(-1)+!w*resid2_B
     equation eq3.ls sempary c x sempary(-1)
     bsimres(!res)=(eq3.@coefs(3)-!beta3)/eq3.@stderrs(3)
   next
   vector rank = @sort(bsimres)
   simres(!rep,2)=simres(!rep,1)<rank(50)
next
simres.write f:\MCS\mcs65sim.txt
```

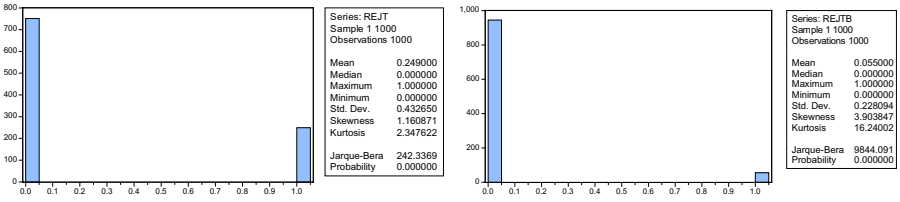

Fig. 6.5 Results from program mcs65.prg.

```
workfile f:\MCS\mcs65sim.wf1 u 1 !R
read f:\MCS\mcs65sim.txt t rejtb
genr rejt=t<@qtdist(0.05,!n-3)
```

The program examines for $n = 40$ and a particular nonstationary exogenous regressor by MCS the type I error probability of the standard asymptotic test (used with critical values from the Student distribution) and the semiparametric bootstrap test of $H_0 : \beta_3 = 0.8$ against $H_1 : \beta_3 < 0.8$ at nominal significance level 5%. Although the disturbances are drawn from the normal distribution, we find very serious size problems for the usual test procedure, whereas the bootstrap, even in this small sample, and not exploiting the normality of the disturbances, cures this problem completely. The MCS estimates of the actual type I error probabilities are 24.9% and 5.5%, respectively. Hence, this is very promising, but deserves of course further study (see exercises).

6.4 Some Further Perspectives

In the first section we did set out to explain techniques by which the qualities of particular econometric inference procedures can be examined by performing well designed experiments on a computer. However, in this final section we ended by discussing how performing experiments can directly contribute to generate inference for particular empirical cases. To put this all into perspective we will first formalize the basic idea underlying the bootstrap, known as the bootstrap principle, and clearly distinguish the parametric and the semiparametric bootstrap.

Let a population be characterized by a CDF given by $F = F_\theta(x)$. Here θ is a vector of parameters which in this particular population has true but unknown value θ^0, and x is a scalar, which can assume

any possible value in the support of F. We have observed $(y_1, \ldots, y_n)$, which is an IID sample of size n from this population, i.e., $y_i \sim F_{\theta^0}$, $i = 1, \ldots, n$. From this empirical sample we can obtain the empirical cumulative distribution function (ECDF) in one of the following two ways. In a pure parametric setting, where F_θ has been fully specified, by first estimating the parameters, giving $\hat{\theta}_n = g(y_1, \ldots, y_n)$, which then yields $\hat{F}_n(x) \equiv F_{\hat{\theta}_n}(x)$. And in a fully nonparametric setting the ECDF can be obtained by $\hat{F}_n(x) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i \leq x)$, where $\mathbb{I}(\cdot)$ is the indicator function.

Now let $\hat{F}_n^*$ denote the ECDF of a sample $(y_1^*, \ldots, y_n^*)$ drawn from the pseudo population characterized by $\hat{F}_n$. From this so-called bootstrap sample $(y_1^*, \ldots, y_n^*)$ we can obtain an estimator $\hat{\theta}_n^* = g(y_1^*, \ldots, y_n^*)$. Hence, in the bootstrap the pair $(F, \hat{F}_n)$ is replaced by the pair $(\hat{F}_n, \hat{F}_n^*)$. Note, though, that in the first pair the true population parameter value θ^0 is unknown, whereas in the second pair its sample equivalent $\hat{\theta}_n$ is known! Therefore, if we are able to assess either analytically or by experiments the major characteristics of statistics which are functions of $(\hat{\theta}_n^* - \hat{\theta}_n)$, then we may use these to approximate the characteristics of the same functions of $(\hat{\theta}_n - \theta^0)$.

When it is hard to derive the major properties of $(\hat{\theta}_n^* - \hat{\theta}_n)$ analytically, as is often the case, then one can exploit the fact that the population is known in the second pair, so that we can simulate drawings from $\hat{F}_n$ and assess the major characteristics of $(\hat{\theta}_n^* - \hat{\theta}_n)$ from simulated realizations $(\hat{\theta}_n^{*(j)} - \hat{\theta}_n)$, $j = 1, \ldots, B$, where B is the sample size of such a bootstrap simulation study.

The above is only worthwhile if the approximation of the characteristics of $(\hat{\theta}_n - \theta^0)$ by those of $(\hat{\theta}_n^* - \hat{\theta}_n)$ involves a refinement of those obtained by standard first-order asymptotic approximations. That is not always the case. However, proofs that bootstrap procedures provide approximations with approximation errors of lower order in n than those of first-order asymptotics have been provided for many particular cases and implementations of the bootstrap simulations.

The primary difference between classic naive Monte Carlo simulation (MCS) and bootstrap simulation (BSS) is that the bootstrap focusses mostly on one particular empirical process, and aims

to provide a direct answer to an actual inference problem faced by a practitioner. The information provided by MCS is more indirect; it can tell how well (or how badly) specific techniques work in particular (often stylized) situations. The practitioner has to make sure that this advice on the qualities of the technique that he is using really concerns the context of his specific empirical problem. In a way BBS can be more useful than MCS because it directly addresses a specific actual empirical inference problem. Nevertheless, BSS cannot replace MCS for the following reason. Apart from a few particular implementations for very special modelling situations, bootstrap inference is only asymptotically valid for n large. In order to analyze the qualities of bootstrap inference in finite samples, and compare it with alternative inference techniques, one again needs MCS and has to design an appropriate grid on a family of Monte Carlo DGPs and all that. There does not seem to be another alternative yet than MCS to obtain evidence on the actual performance of the bootstrap in finite samples for particular DGPs and inference techniques. Note that such MCS studies are extremely computer intensive, because all separate R Monte Carlo replications will then require B bootstrap replications each.

Exercises

1. How would the formulas (4.5), (4.6), and (4.7) be affected if the test procedure is such that small values of q_n should lead to rejection of H_0?
2. Is the Durbin–Watson bounds test when performed in the classic normal linear regression model an exact test? If one would make the inconclusive region part of the critical region, would that make the test conservative (or liberal)? Would that increase or decrease the probability of type II errors?
3. Is it possible to perform an exact test based on the Durbin–Watson test in the ARX(1) model, by using it as a Monte Carlo test? Why (not)?
4. Adapt program mcs61.prg such that you can perform an exact Monte Carlo test for first-order serial correlation in the classic regression model with fat tailed disturbances, such

that under H_0 one has $u_i \sim IID(0,\sigma^2)$ with $u_i = \sigma v_i/\sqrt{3}$ and v_i student distributed with 3 degrees of freedom.

5. Although the Help files of EViews suggest that it uses the JB formula as given in (6.6), it can be checked that this is not really the case, since both JB (and the s that it uses) are obtained in EViews by taking n instead of $n-k$. Explain what the effects of that will be for the asymptotic properties of the test, and may be for the finite sample properties of the test procedure. From this, draw a more general conclusion on the relevance for PMC test procedures of degrees of freedom corrections and other transformations aimed at improving the finite sample properties of a test.
6. Adapt the programs mcs63.prg and mcs64.prg such that you test $H_0 : \beta_2 + \beta_3 = 0$ against one-sided alternatives both by a standard asymptotic test and its bootstrap alternative.
7. Adapt programs mcs63.prg and mcs64.prg such that you test $H_0 : \gamma = -0.4$ against $H_1 : \gamma > -0.4$.
8. Adapt program mcs65.prg such that you get a better picture of the performance of the two test procedures. Make changes as follows: Increase the value of R such that the MCS results have an accuracy that you find acceptable (argue your choice). Examine how well the bootstrap works for much smaller values of B than 999, say 199. Examine rejection probabilities against both left-hand ($\beta_3 < \beta_{30}$) and right-hand ($\beta_3 > \beta_{30}$) alternatives. Stick to conditioning on just one realization of the x_i series. It may be a good idea (to reduce computing time) to reduce the value of n to say 20. Examine the effect of σ, and also of the true value of β_3.

A

Appendices

A.1 Major Issues in Classic Econometric Inference

1. *Means*: Econometric models specify parametric relationships between observed variables allowing for unexplained aspects through unobserved random disturbances. Estimation and testing techniques, which exploit probabilistic assumptions on the disturbances, are employed in order to produce inference on the unknown parameter values in the form of point estimates, their estimated standard deviations, supplemented by test statistics and their p-values and associated confidence sets.

2. *Goals*: These inferences are used to corroborate the tenability of the various assumptions made, given the empirical evidence obtained from the data. And, if these assumptions seem to hold, to verify or quantify economic theory regarding the numerical values of reaction coefficient parameters. This could next form a stepping stone toward forecasting future values of endogenous variables, and also for policy analysis, in which by manipulating the values of particular exogenous variables aimed at goals set for the endogenous variables might be achieved.

3. *Quality*: The actual accuracy of the inference obtained by these techniques depends on the adequacy of the model specification and the other model assumptions made regarding the properties of the probability distribution of the disturbance terms, so on the match between the adopted probabilistic characteristics of the chosen model and the true properties of the underlying actual data generating process (DGP).

4. *Problems*: In empirically relevant situations it is usually the case that neither the central tendency nor the dispersion of econometric estimators of the model parameters can be derived analytically. So, usually precise measures on the accuracy of parameter estimates are not available. In fact, estimators of parameters and estimators of the variance of parameter estimates will usually be biased, and the assumed distribution of test statistics under the null hypothesis will just be an approximation, and possibly a poor one. Therefore, actual significance levels of tests and the actual coverage probability of confidence regions will be largely unknown.

5. *Warning*: Only under very precise and strict (and therefore usually unrealistic) assumptions the ideal of a best unbiased estimator is achievable. And in order to be able to obtain a test statistic for testing a parametric hypothesis at a pre-specified nominal significance level (maximum allowed probability of committing a type I error) we have to be even more specific, because these would require a firm and valid assumption on the (normality of the) distribution function of the random disturbances and exogeneity of all regressors. This may occur in laboratory experiments but is highly unlikely for observational data.

6. *Distribution function*: Given a particular empirical data set of sample size n and some chosen model specification involving parameters $\theta = (\beta, \dots)$, where the vector β contains the parameters of primary interest, an estimation technique (OLS, GLS, FGLS, IV, NLS, MLE, GMM, FGMM, CUE, SURE, 2SLS, 3SLS, LIML, FIML, ...) yields just one point-estimate $\hat{\beta}$, obtained from one specific sample realization. To produce further statistical inference on the true numerical value β^0 of β in the form of hypotheses tests or confidence regions we need to know as much as possible about the stochastic properties of $\hat{\beta}$, preferably its probability distribution (CDF/PDF) over its sample-population.

7. *Moments*: Only in very specific cases the assumptions made on the random elements of the model are such that the distribution function of $\hat{\beta}$ and its first two moments

$$E(\hat{\beta}) \text{ and } Var(\hat{\beta}) \equiv E\{[(\hat{\beta} - E(\hat{\beta})][(\hat{\beta} - E(\hat{\beta})]'\} \tag{A.1}$$

can be derived analytically (possibly conditional on realized exogenous variables). In such rare cases the bias $E(\hat{\beta}) - \beta^0$ and the mean squared error $MSE(\hat{\beta}) \equiv E[(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)']$, which will depend on the data and θ^0, can be calculated for values within the relevant parameter space, and be compared with those of alternative estimation procedures.

8. *Finite sample tests*: Only in even more specific situations exact tests (these reject a true null hypothesis with a probability that never exceeds the chosen nominal significance level) can be performed and corresponding exact confidence sets constructed (these have a probability to cover the true parameter values by a probability not smaller than the nominal confidence coefficient). Only then proper comparisons can be made between the power (probability to reject a false hypothesis, given a chosen maximum probability to reject it when true) of alternative tests and between the width/volume of corresponding confidence sets.

9. *Asymptotic alibi*: Usually, analytical results cannot be obtained in finite samples, because estimators are mostly highly nonlinear functions of various random variables. Then the distributions of such estimators and of related test statistics (also when the tested null hypothesis is true) depend in a complicated and often intractable way on the values of so-called nuisance parameters. However, under additional so-called regularity assumptions, approximations to the finite sample distribution can be obtained by using (standard) large sample asymptotic methods for $n \to \infty$.

10. *Asymptotic properties*: Under (preferably mild) regularity conditions, a law of large numbers can be invoked that may yield that the estimator $\hat{\beta}$ is consistent, i.e.,

$$\text{plim}_{n\to\infty}\hat{\beta} = \beta^0. \tag{A.2}$$

And next, possibly, a central limit theorem can be employed, yielding the limiting normal distribution

$$n^{1/2}(\hat{\beta} - \beta^0) \xrightarrow{d} N[0, V^{\infty}(\hat{\beta})], \tag{A.3}$$

for $n \to \infty$, where $V^{\infty}(\hat{\beta})$ should be finite. If for all alternative consistent estimators $\hat{\beta}^*$, having a limiting distribution with asymptotic variance $V^{\infty}(\hat{\beta}^*)$, the matrix difference $V^{\infty}(\hat{\beta}^*) - V^{\infty}(\hat{\beta})$ is positive semi definite, then estimator $\hat{\beta}$ is asymptotically efficient.

11. *Asymptotic tests*: To produce inference on β based on $\hat{\beta}$ and its limiting normal distribution, one should be able to estimate $V^{\infty}(\hat{\beta})$ consistently too, say by $\hat{V}^{\infty}(\hat{\beta})$. The resulting asymptotic tests are exact only for $n \to \infty$ and therefore in finite samples their actual significance level (probability to reject a true null) may deviate from the chosen nominal significance level. This may be clarified as follows. Consider the standard asymptotic Wald-type test for $H_0 : \beta_j = \beta_{j0}$ against the one-sided alternative $H_1 : \beta_j > \beta_{j0}$, where β_j is an arbitrary scalar element from β, and β_{j0} is a hypothesized real number for the true value β_j^0. The Wald-type test statistic for this null-hypothesis is given by

$$W_j = (\hat{\beta}_j - \beta_{j0})/[\hat{V}^{\infty}(\hat{\beta}_j)/n]^{1/2}. \tag{A.4}$$

When H_0 is true (i.e., $\beta_{j0} = \beta_j^0$) then $W_j \underset{n\to\infty}{\to} N(0,1)$. If z_α denotes the α-th quantile of the standard normal distribution, hence $(2\pi)^{-1/2} \int_{-\infty}^{z_\alpha} \exp(-z^2/2) dz = \alpha$, then the actual rejection probability of this asymptotic test against this (right-hand-side) alternative at nominal significance level α equals in finite samples

$$r_j(\alpha; n, \theta, \beta_{j0}) \equiv \Pr\{W_j > z_{1-\alpha}\}. \tag{A.5}$$

When evaluated for $\beta_{j0} = \beta_j^0$ this probability is the actual significance level of the test. In finite samples this will usually differ from α and vary with θ and n. This even occurs in the classic normal linear regression model, but can be avoided by not using the asymptotic critical values $z_{1-\alpha}$ of $N(0,1)$ but the quantiles of a Student distribution (which correspond to those of the standard normal only when n is large).

12. *Type I and II error balance*: So, the actual type I error probability will usually differ in finite samples from the nominal significance level α.

In that case for any $\beta_{j0} \neq \beta_j^0$ the rejection probability $r_j(\alpha; n, \theta, \beta_{j0})$ does no longer adequately represent the power of the test, because of the following. Imagine that for $\alpha = 0.05$ and some particular θ with $\beta_{j0} \neq \beta_j^0$ we have $r_j(\alpha; n, \theta, \beta_{j0}) = 0.8$. This rejection probability of a false null would be reasonably attractive if for all θ with $\beta_j = \beta_j^0$ probability $r_j(\alpha; n, \theta, \beta_j^0) = 0.05$, but not when it could occur that $r_j(\alpha; n, \theta, \beta_j^0) = 0.7$.

13. *Size problems*: The size of a test is the maximum of its rejection probability over the whole subset of the parameter space that respects the null hypothesis. Hence,

$$size(\alpha; n, \beta_{j0}) = \max_{\theta,\ \beta_{j0}=\beta_j^0} \{r_j(\alpha; n, \theta, \beta_{j0})\}, \tag{A.6}$$

where the maximum is taken by keeping α, n and $\beta_{j0} = \beta_j^0$ fixed and varying all other elements of the parameters $\theta = (\beta, \dots)$. Note that these other parameters are genuinely nuisance parameters when $r_j(\alpha; n, \theta, \beta_{j0})$ is not invariant with respect to them. If in a particular context the size is either smaller than $(1 - \phi)\alpha$ or larger than $(1 + \phi)\alpha$ then there are serious size problems if, say, $\phi > 1/3$.

14. *Assessing finite sample properties*: Because large-sample asymptotic approximations may be inaccurate in finite samples (consistent estimators may have a huge bias, the actual variance $\mathrm{Var}(\hat{\beta})$ may differ substantially from $\hat{V}^{\infty}(\hat{\beta})/n$, an asymptotically efficient estimator may have a larger mean squared error in finite samples than an inconsistent estimator, and the actual type I error probability of a test may differ seriously from the nominal significance level, etc.) such asymptotic approximations may be less useful for rating techniques (estimators and tests) regarding their properties in finite samples. To analyze and compare the performance in finite samples of competing techniques there are basically two options. One is analytical and the other experimental.

15. *Higher-order asymptotics*: Ideally one would like to derive precise analytic finite sample results on the behavior of estimators and tests, but very often these are not tractable. Then, as a second best approach, one may aim to obtain better analytic approximations by using asymptotic expansions yielding higher-order asymptotic

approximations. This employs a decomposition of $n^{1/2}(\hat{\beta} - \beta^0)$ of the form of

$$n^{1/2}(\hat{\beta} - \beta^0) = \xi_1 + n^{-1/2}\xi_2 + n^{-1}\xi_3 + \cdots, \tag{A.7}$$

where the ξ_i, $i \geq 1$ have "a finite distribution", which means that for any real $\varepsilon > 0$ there is a constant $c < \infty$ such that $\Pr(|v_i| > c) < \varepsilon$ for any n (hence, the probability mass for very large values is negligible). Note that the standard asymptotic procedures provide a first-order approximation and just focus on ξ_1, which in many regular cases has $\xi_1 \sim N[0, V^{\infty}(\hat{\beta})]$. By taking higher-order terms into account as well, one may obtain more accurate approximations, either to the distribution of $n^{1/2}(\hat{\beta} - \beta^0)$ or just to its moments.

16. *Monte Carlo simulation*: Experimental evidence on the actual characteristics of the finite sample distribution of estimators and tests for relevant parameter values can always be obtained, avoiding the complexities of analytical derivations, from so-called Monte Carlo experiments. These too involve particular approximations and thus give rise to inaccuracies. Moreover, their results are usually very specific for the chosen experimental design, hence despite their potential they have serious limitations too.

17. *Limitations of simulations*: Monte Carlo experiments are at best asymptotic approximations themselves. However, they are not asymptotic regarding the size of the (econometric) sample n, but regarding the size of the sample of experiments performed in the Monte Carlo simulation, say R. Since we can choose R ourselves, and we may choose it rather large (depending on the computer speed and facilities available) the asymptotic approximations involved in a Monte Carlo analysis can be made very accurate, and the remaining approximation errors can be controlled and checked very well, although in practice it is not done very often. However, the results obtained by standard Monte Carlo experiments are always very specific since they are based on a full characterization of the (in practice unknown) DGP.

18. *A further perspective*: Simulations can be made less dependent on the specified but in practice unknown parameter values and be geared

toward the actual empirical inference problem at hand by resampling the available actual data and generating so-called bootstrap inference. In particular cases this provides a higher-order asymptotic approximation by experimentation without the need to explicitly derive the analytic results.

B

Tools for Standard Asymptotic Analysis

To support the understanding of the large-R asymptotic foundations of Monte Carlo analysis, as well as the large-n asymptotic analysis of econometric estimation techniques and test procedures, we here summarize some basic essential tools for asymptotic derivations. We do so in terms of the familiar n.

B.1 Big O and Little o.

If $\{a_n\}$ is a scalar non-stochastic sequence of real numbers for integer $n = n_0, \ldots, \infty$ then

$$\begin{array}{ll} a_n = O(1), & \text{if } |a_n| < c, \quad \forall n \quad \text{and} \quad 0 < c < \infty; \\ a_n = O(n^h), & \text{if } n^{-h} a_n = O(1); \\ a_n = o(n^h), & \text{if } \lim_{n \to \infty} n^{-h} a_n = 0. \end{array}$$

Hence, $a_n = O(1)$ means that a_n is bounded (i.e., is at most of order $1 = n^0$) and $a_n = O(n^h)$ indicates that a_n is at most of order n^h, whereas $a_n = o(n^h)$ implies that a_n is of smaller order than n^h.

For $\{b_n\}$ a scalar sequence of random variables we say that b_n is of order n^g in probability, and express that as $b_n = O_p(n^g)$, if for any

$\varepsilon > 0$ there is a constant $c < \infty$ such that

$$\Pr(|n^{-g} b_n| > c) < \varepsilon \quad \text{for all } n > n_0.$$

If $b_n = O_p(n^g)$ and $d_n = O_p(n^h)$ or $d_n = O(n^h)$ then

$$b_n d_n = O_p(n^{g+h}) \quad \text{and} \quad b_n + d_n = O_p(n^{\max(g,h)}).$$

Note that if $b_n = O_p(n^g)$ then $b_n = o(n^h)$, for all $h > g$. Also, $b_n = O_p(n^g)$ implies $b_n^{-1} = O_p(n^{-g})$. For a vector (or matrix) being $O_p(n^g)$ means that all its separate elements are $O_p(n^g)$.

B.2 Important Laws, Lemmas and Theorems

Let the $k \times 1$ vector v_i for $i = 1, \ldots, n$ be such that $E(v_i) = \mu_i = O(1)$ and $\mathrm{Var}(v_i) = O(1)$ (first two moments of all v_i exist and are finite). Below we shall consider $\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i$ and $\bar{\mu} = \frac{1}{n}\sum_{i=1}^n \mu_i$, but also more general (usually k element) functions $g_n = g(v_1, \ldots, v_n)$ and functions (often with fewer than k elements) $h(g_n)$.

Law of large numbers (LLN):
If $\lim_{n\to\infty} \mathrm{Var}(\bar{v}) = O$ (note that this O is a matrix filled with zero elements), which is certainly the case if the v_i are mutually uncorrelated but also if they are asymptotically uncorrelated, then $\mathrm{plim}_{n\to\infty} \bar{v} = \lim_{n\to\infty} \bar{\mu}$. Weaker versions exist which do not require the existence of second moments.

Corollary for finding probability limits:
If $\lim_{n\to\infty} E(g_n) = O(1)$ and $\lim_{n\to\infty} \mathrm{Var}(g_n) = o(1)$ then $\mathrm{plim}_{n\to\infty} g_n = \lim_{n\to\infty} E(g_n)$.

Central limit theorem (CLT):
If $\mathrm{Cov}(v_i, v_j) = O$ (again a matrix of zeros) for $i \neq j$ then

$$n^{1/2}(\bar{v} - \bar{\mu}) \xrightarrow[n\to\infty]{d} N\left(0, \lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathrm{Var}(v_i)\right).$$

Slutsky:
Let $h(\cdot)$ be a continuous function not involving n then $\mathrm{plim}_{n\to\infty} h(g_n) = h(\mathrm{plim}_{n\to\infty} g_n)$. Note that $h(g_n)$ could for v_i scalar for instance be $\bar{v}^2$ or $\bar{v}^{-1}$.

Cramér:
Let $\hat{\Omega}_n$ and Ω be $k \times k$ matrices, A_n and A matrices of order $l \times k$ and γ a $k \times 1$ vector. Also let $\text{plim}_{n\to\infty} g_n = \gamma$, $\text{plim}_{n\to\infty} \hat{\Omega}_n = \Omega$, $\text{plim}_{n\to\infty} A_n = A$ with γ, Ω and A deterministic, and $n^{1/2}(g_n - \gamma) \underset{n\to\infty}{\overset{d}{\to}} \tilde{g} \sim N(0,\Omega)$. Then $n^{1/2} A_n(g_n - \gamma) \underset{n\to\infty}{\overset{d}{\to}} \tilde{g}^* \sim N(0, A\Omega A')$ and the first-order asymptotic approximation for the distribution of $A_n g_n$ is $N(A_n\gamma, n^{-1} A_n \hat{\Omega}_n A_n')$.

Continuous mapping theorem (CMT):
Let $g_n \underset{n\to\infty}{\overset{d}{\to}} g$, hence g is the limiting distribution of g_n, then $h(g_n) \underset{n\to\infty}{\overset{d}{\to}} h(g)$.

Delta-method:
Let $n^{1/2}(g_n - \gamma) \underset{n\to\infty}{\overset{d}{\to}} \tilde{g} \sim N(0,\Omega)$ and $h(\cdot)$ be a differentiable vector function not involving n then $n^{1/2}[h(g_n) - h(\gamma)] \underset{n\to\infty}{\overset{d}{\to}} \tilde{h} \sim N(0, (\frac{\partial h(g)}{\partial g'})\Omega(\frac{\partial h(g)}{\partial g'})')$, where $(\frac{\partial h(g)}{\partial g'})_{ij} \equiv \frac{\partial h_i(g)}{\partial g_j}$.

Exercises

1. Show that $\text{MSE}(\hat{\beta}) = \text{Var}(\hat{\beta}) + B(\hat{\beta})B(\hat{\beta})'$, where $B(\hat{\beta}) \equiv E(\hat{\beta}) - \beta^0$ denotes the bias vector.
2. Let estimator $\hat{\beta}$ be such that $n^{1/2}(\hat{\beta} - \beta^0) \overset{d}{\to} N[0, V^\infty(\hat{\beta})]$, with $V^\infty(\hat{\beta}) = O(1)$. Discuss what this does (not?) imply for $E(\hat{\beta})$, for $Var(\hat{\beta})$, for the form of the distribution of $\hat{\beta}$, and for $\text{plim}\,\hat{\beta}$. Is it possible to say something about the limiting distribution of $\hat{\beta}$, or of $n^{1/2}\hat{\beta}$?
3. Let $\text{plim}\,\hat{V}^\infty(\hat{\beta}) = V^\infty(\hat{\beta})$. Then we use in finite samples $\hat{V}(\hat{\beta}) = \hat{V}^\infty(\hat{\beta})/n$ to estimate $\text{Var}(\hat{\beta})$. Why? Show that $\text{plim}\,\hat{V}(\hat{\beta}) = O$, whereas it seems doubtful that $\text{Var}(\hat{\beta})$ is O in finite samples.
4. Give concise definitions of and state the major characteristics of inference issues in econometrics, giving special attention to: bias, consistency, limiting distribution, efficiency, small sample distribution, nominal significance level, actual signif-

icance level, test size, test power, actual null distribution of a test, nuisance parameters.

5. Consider the classic linear regression model $y = X\beta + u$, where the k regressors in X are **exogenous** and the disturbances $u_i \mid X \sim IID(0, \sigma^2)$ for $i = 1, \ldots, n$. Consider the two cases (a) $u_i \sim NIID(0, \sigma^2)$ and (b) $u_i = \sigma(\eta_i - d)/\sqrt{2d}$ where $\eta_i \sim \chi^2_d$. Note that in case (a) the disturbances are symmetric, and in case (b) they are skew to the right. Let $t^{1-\alpha}_{n-k}$ represent the $(1-\alpha)^{th}$ quantile of the Student distribution with $n-k$ degrees of freedom and consider the Student test statistic for $H_0 : \beta_j = \beta_{j0}$ given by

$$T_j = \frac{\hat{\beta}_j - \beta_{j0}}{s\sqrt{[(X'X)^{-1}]_{jj}}},$$

where $s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(n-k)$. Discuss the possible effects of nuisance parameters on the null-distribution of this test in the cases (a) and (b), respectively.

6. In the context of the standard linear regression model $y = X\beta + u$, where $X = (x_1, \ldots, x_n)'$ and $u = (u_1, \ldots, u_n)'$ with $u_i \mid x_i \sim IID(0, \sigma^2)$, we obtain asymptotic results for OLS by considering $v_i = x_i u_i$. Exploit the theorems given above and state the required further regularity in order to establish consistency and the limiting distribution of the OLS estimator

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= \beta + (X'X)^{-1}X'u = \beta + \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} v_i.\end{aligned}$$

7. In the model of the former exercise, examine what the characteristics of the regressor variables have to be in order to have $(X'X)^{-1} = O_p(n^{-1})$ and $X'u = O_p(n^{-1/2})$. What are the implications for the order of probability of $\hat{\beta} - \beta_0$?

References

Arellano, M. and S. R. Bond (1991), 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations'. *The Review of Economic Studies* **58**, 277–297.

Brown, B. W. and W. K. Newey (2002), 'Generalized method of moments, efficient bootstrapping, and improved inference'. *Journal of Business and Economic Statistics* **20**, 507–517.

Bun, M. J. G. and J. F. Kiviet (2006), 'The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models'. *Journal of Econometrics* **132**, 409–444.

Cramér, H. (1946), *Mathematical Methods of Statistics*. Princeton.

David, H. A. (1981), *Order Statistics*. Wiley.

Davidson, R. and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Davidson, R. and J. G. MacKinnon (1998), 'Graphical methods for investigating the size and power of hypothesis tests'. *The Manchester School* **66**, 1–26.

Davidson, R. and J. G. MacKinnon (2004), *Econometric Theory and Methods*. Oxford University Press.

Davidson, R. and J. G. MacKinnon (2006), 'Bootstrap methods in econometrics'. In: T. C. Mills and K. Patterson (eds.): *Palgrave*

Handbook of Econometrics, vol. 1*, Econometric Theory*. Basingstoke, Palgrave Macmillan, pp. 812–838.

Doornik, J. A. (2006), 'The role of simulation in econometrics'. In: T. C. Mills and K. Patterson (eds.): *Palgrave Handbooks of Econometrics*, vol. 1*, Econometric Theory*. Basingstoke, Palgrave Macmillan, pp. 787–811.

Dufour, J.-M. and L. Khalaf (2001), 'Monte Carlo test methods in econometrics'. In: B. Baltagi (ed.): *Companion to Theoretical Econometrics*. Oxford, UK: Blackwell, pp. 494–519.

Edgerton, D. L. (1996), 'Should stochastic or non-stochastic exogenous variables be used in Monte Carlo experiments?'. *Economics Letters* **53**, 153–159.

Fishman, G. S. (2006), *A First Course in Monte Carlo*. Thomson.

Gourieroux, C. and A. Monfort (1996), *Simulation-Based Econometric Methods*. Oxford University Press.

Greene, W. H. (2012), *Econometric Analysis*. 7th international edition. Pearson.

Hendry, D. F. (1984), 'Monte Carlo experimentation in econometrics'. In: Z. Griliches and M. D. Intriligator (eds.): *Handbook of Econometrics*, vol. II. Amsterdam: Elsevier.

Hendry, D. F. (1995), *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D. F. and B. Nielsen (2007), *Econometric Modelling*. Oxford: Princeton University Press.

Horowitz, J. L. (2003), 'The bootstrap in econometrics'. *Statistical Science* **18**, 211–218.

Johnson, R. W. (2001), 'An introduction to the bootstrap'. *Teaching Statistics* **23**, 49–54.

Kiviet, J. F. (1985), 'Model selection test procedures in a single linear equation of a dynamic simultaneous system and their defects in small samples'. *Journal of Econometrics* **28**, 327–362.

Kiviet, J. F. (2007), 'Judging contending estimators by simulation: Tournaments in dynamic panel data models'. In: G. D. A. Phillips and E. Tzavalis (eds.): *The Refinement of Econometric Estimation and Test Procedures; Finite Sample and Asymptotic Analysis*. Cambridge University Press, pp. 282–318.

Kiviet, J. F. and J. Niemczyk (2011), 'Comparing the asymptotic and empirical (un)conditional distributions of OLS and IV in a linear static simultaneous equation'. To appear in*: Journal of Computational Statistics & Data Analysis*.

Kiviet, J. F. and G. D. A. Phillips (2012), 'Higher-order asymptotic expansions of the least-squares estimation bias in first-order dynamic regression models'. To appear in: *Journal of Computational Statistics & Data Analysis*.

MacKinnon, J. G. (2002), 'Bootstrap inference in econometrics'. *Canadian Journal of Economics* **35**, 615–645.

MacKinnon, J. G. (2006), 'Bootstrap methods in econometrics'. *Economic Record* **82**, s2–s18.

Metropolis, N. and S. Ulam (1949), 'The Monte Carlo method'. *Journal of the American Statistical Society* **44**, 335–341.

Nelson, R. N. and R. Startz (1990), 'Some further results on the exact small sample properties of the instrumental variable estimator'. *Econometrica* **58**, 967–976.

Stigler, S. M. (1999), *Statistics on the Table; The History of Statistical Concepts and Methods*. Cambridge MA: Harvard University Press.

Student (1908), 'On the probable error of the mean'. *Biometrika* **6**, 1–25.

Wagner, H. M. (1958), 'A Monte Carlo study of estimates of simultaneous linear structural equations'. *Econometrica* **26**, 117–133.