Bootstrap Methods in Econometrics

by

Russell Davidson

Department of Economics
McGill University
Montreal, Quebec, Canada
H3A 2T7

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

email: **russell.davidson@mcgill.ca**

and

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

email: **jgm@econ.queensu.ca**

Abstract

Although it is common to refer to "the bootstrap," there are actually a great many different bootstrap methods that can be used in econometrics. We emphasize the use of bootstrap methods for inference, particularly hypothesis testing, and we also discuss bootstrap confidence intervals. There are important cases in which bootstrap inference tends to be more accurate than asymptotic inference. However, it is not always easy to generate bootstrap samples in a way that makes bootstrap inference even asymptotically valid.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada.

October, 2004

1. Introduction

When we perform a hypothesis test in econometrics, we reject the null hypothesis if the test statistic is unlikely to have occurred by chance for the distribution that it should follow under the null. Traditionally, this distribution is obtained by theoretical methods. In many cases, we use asymptotic distributions that are strictly valid only if the sample size is infinitely large. In others, such as the classical normal linear regression model with its associated t and F statistics, we use finite-sample distributions that depend on very strong distributional assumptions.

Bootstrap methods, which have become increasingly popular in econometrics during the last decade as the cost of computation has fallen dramatically, provide an alternative way of obtaining the distributions to which test statistics are to be compared. The idea is to generate a large number of simulated **bootstrap samples**, use each of them to calculate a **bootstrap test statistic**, and then compare the actual test statistic with the empirical distribution of the bootstrap statistics. When the latter provides a good approximation to the unknown true distribution of the test statistic under the null hypothesis, bootstrap tests should lead to accurate inferences. In some cases, they lead to very much more accurate inferences than using asymptotic distributions in the traditional way.

Because of the close connection between hypothesis tests and confidence intervals, we can also obtain **bootstrap confidence intervals**. The ends of a confidence interval generally depend on the quantiles of the distribution that some test statistic is supposed to follow. We "invert" the test to find the interval that contains all the parameter values that would not be rejected by the test. One way to construct a bootstrap confidence interval is to use the quantiles of the empirical distribution of a set of bootstrap test statistics instead of the quantiles of a theoretical distribution.

In the next section, we discuss the basic ideas of bootstrap testing for the special case in which the bootstrap statistics follow exactly the same distribution as the actual test statistic under the null hypothesis. We show that, in this special case, it is possible to perform a **Monte Carlo test** that is exact, in the sense that the actual probability of Type I error is equal to the nominal significance level of the test. This is an important result, in part because there are valuable applications of Monte Carlo tests in econometrics, and in part because this case serves as a benchmark for bootstrap tests more generally.

In most cases, it is impossible to find a **bootstrap data generating process**, or **bootstrap DGP**, such that the bootstrap statistics follow exactly the same distribution as the actual test statistic under the null hypothesis. In Section 3, we discuss the **parametric bootstrap**, for which the bootstrap DGP is completely characterized by a set of parameters that can be consistently estimated. We show that, under mild regularity conditions likely to be satisfied in many cases of interest in econometrics, the error in rejection probability (ERP) of a parametric bootstrap test, that is, the difference between the true probability of rejecting a true null hypothesis and the

nominal level, tends to zero as the sample size tends to infinity faster than does the ERP of conventional asymptotic tests.

In Sections 4-8, we discuss various methods for the construction of bootstrap DGPs that are applicable to a variety of problems in econometrics. In our view, this is where the principal impediments to the widespread adoption of bootstrap methods lie. Methods that work well for some problems may be invalid or may perform poorly for others. Econometricians face a great many challenges in devising bootstrap DGPs that will lead to accurate inferences for many of the models that we commonly estimate.

In Section 9, we extend the methods of bootstrap testing discussed earlier in the paper to the construction of confidence intervals. Finally, Section 10 contains a general discussion of the accuracy of bootstrap methods and some concluding remarks.

2. Monte Carlo Tests

The simplest type of bootstrap test, and the only type that can be exact in finite samples, is called a **Monte Carlo test**. This type of test was first proposed by Dwass (1957). Monte Carlo tests are available whenever a test statistic is **pivotal**. Let τ denote a statistic intended to test a given null hypothesis. By **hypothesis** we mean a set of DGPs that satisfy some condition or conditions that we wish to test. Then the statistic τ is pivotal for this null hypothesis if and only if, for each possible fixed sample size, the distribution of τ is the same for all of the DGPs that satisfy the hypothesis. Such a test statistic is said to be a **pivot**.

Suppose we compute a realization $\hat{\tau}$ of a pivotal test statistic using real data, and then compute B independent bootstrap test statistics τ_j^*, $j = 1, \ldots, B$, using data simulated using any DGP that satisfies the null hypothesis. Since τ is a pivot, it follows that the τ_j^* and $\hat{\tau}$ are independent drawings from one and the same distribution, *provided* that the true DGP, the one that generated $\hat{\tau}$, also satisfies the null hypothesis.

Imagine that we wish to perform a test at significance level α, where α might, for example, be .05 or .01, and reject the null hypothesis when the value of $\hat{\tau}$ is unusually large. Given the actual and simulated test statistics, we can compute a **bootstrap *P* value** as

$$p^*(\hat{\tau}) = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_j^* > \hat{\tau}), \tag{1}$$

where $\mathrm{I}(\cdot)$ is the **indicator function**, with value 1 when its argument is true and 0 otherwise. Evidently, $p^*(\hat{\tau})$ is just the fraction of the bootstrap samples for which τ_j^* is larger than $\hat{\tau}$. If this fraction is smaller than α, we reject the null hypothesis. This makes sense, since $\hat{\tau}$ is extreme relative to the empirical distribution of the τ_j^* when $p^*(\hat{\tau})$ is small.

Now suppose that we sort the original test statistic $\hat{\tau}$ and the B bootstrap statistics τ_j^* in decreasing order. Define the rank r of $\hat{\tau}$ in the sorted set in such a way that there

are exactly r simulations for which $\tau_j^* > \hat{\tau}$. Then r can have $B+1$ possible values, $r = 0, 1, \ldots, B$, all of them equally likely under the null. The estimated P value $\hat{p}^*(\hat{\tau})$ is then just r/B.

The Monte Carlo test rejects if $r/B < \alpha$, that is, if $r < \alpha B$. Under the null, the probability that this inequality is satisfied is the proportion of the $B+1$ possible values of r that satisfy it. If we denote by $[\alpha B]$ the largest integer that is smaller than αB, there are exactly $[\alpha B]+1$ such values of r, namely, $0, 1, \ldots, [\alpha B]$. Thus the probability of rejection is $([\alpha B]+1)/(B+1)$. We want this probability to be exactly equal to α. For that to be true, we require that

$$\alpha(B+1) = [\alpha B] + 1.$$

Since the right-hand side above is the sum of two integers, this equality can hold only if $\alpha(B+1)$ is also an integer. In fact, it is easy to see that the equation holds whenever $\alpha(B+1)$ is an integer. In that case, therefore, the rejection probability under the null, that is, the Type I error of the test, is precisely α, the desired significance level.

Of course, using simulation injects randomness into this test procedure, and the cost of this randomness is a loss of power. A test based on $B = 99$ simulations will be less powerful than a test based on $B = 199$, which in turn will be less powerful than one based on $B = 299$, and so on; see Jöckel (1986) and Davidson and MacKinnon (2000). Notice that all of these values of B have the property that $\alpha(B+1)$ is an integer whenever α is an integer percentage like .01, .05, or .10.

For an example of a Monte Carlo test, consider the classical normal linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \tag{2}$$

where there are n observations, $\boldsymbol{\beta}$ is a k–vector, and the $1 \times k$ vector of regressors $\boldsymbol{X}_t$, which is the t^{th} row of the $n \times k$ matrix $\boldsymbol{X}$, is treated as fixed. Every DGP belonging to this model is completely characterized by the values of the parameter vector $\boldsymbol{\beta}$ and the variance σ^2. Thus any test statistic the distribution of which does not depend on these values is a pivot for the hypothesis that (2) is correctly specified. In particular, a statistic that depends on $\boldsymbol{y}$ only through the OLS residuals and is invariant to the scale of $\boldsymbol{y}$ is pivotal. To see this, note that the vector of OLS residuals is $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, where $\boldsymbol{M}_{\boldsymbol{X}}$ is the orthogonal projection matrix $\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Thus $\hat{\boldsymbol{u}}$ is unchanged if the value of $\boldsymbol{\beta}$ changes, and a change in the variance σ^2 changes $\hat{\boldsymbol{u}}$ only by a scale factor of σ.

One such pivotal test statistic is the estimated autoregressive parameter $\hat{\rho}$ that is obtained by regressing the t^{th} residual $\hat{u}_t$ on its predecessor $\hat{u}_{t-1}$. The estimate $\hat{\rho}$ can be used as a test for serial correlation of the error terms in (2). Evidently,

$$\hat{\rho} = \frac{\sum_{t=2}^n \hat{u}_{t-1}\hat{u}_t}{\sum_{t=2}^n \hat{u}_{t-1}^2}. \tag{3}$$

Since $\hat{u}_t$ is proportional to σ, there are implicitly two factors of σ in the numerator and two in the denominator of (3). Thus $\hat{\rho}$ is independent of the scale factor σ.

Of course, the distribution of $\hat{\rho}$ *does* depend on the regressor matrix $\boldsymbol{X}$. But recall that we assumed that $\boldsymbol{X}$ is fixed in the definition of the model (2). This means that it is the same for every DGP in the model. With this definition, then, $\hat{\rho}$ is indeed pivotal. If we are unwilling to assume that $\boldsymbol{X}$ is a fixed matrix, then we can interpret (2) as a model defined conditional on $\boldsymbol{X}$, in which case $\hat{\rho}$ is pivotal conditional on $\boldsymbol{X}$.

The fact that $\hat{\rho}$ is a pivot means that we can perform an exact Monte Carlo test of the hypothesis that $\rho = 0$ without knowing the distribution of $\hat{\rho}$. The bootstrap DGP used to generate simulated samples can be any DGP in the model (2), and so we may choose the simplest such model, which has $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma^2 = 1$. This bootstrap DGP can be written as

$$y_t^* = \varepsilon_t^*, \quad \varepsilon_t^* \sim \text{NID}(0, 1).$$

For each of B bootstrap samples, we then proceed as follows:

1. Generate the vector $\boldsymbol{y}^*$ as an n–vector of IID standard normal variables.
2. Regress $\boldsymbol{y}^*$ on $\boldsymbol{X}$ and save the vector of residuals $\hat{\boldsymbol{u}}^*$.
3. Compute ρ^* by regressing $\hat{u}_t^*$ on $\hat{u}_{t-1}^*$ for observations 2 through n.

Denote by ρ_j^*, $j = 1, \ldots, B$, the bootstrap statistics obtained by performing the above three steps B times. We now have to choose the alternative to our null hypothesis of no serial correlation. If the alternative is positive serial correlation, then, analogously to (1), we perform a one-tailed test by computing the bootstrap P value as

$$\hat{p}^*(\hat{\rho}) = \frac{1}{B} \sum_{j=1}^{B} \text{I}(\rho_j^* > \hat{\rho}).$$

This P value is small when $\hat{\rho}$ is positive and sufficiently large, thereby indicating positive serial correlation. However, we may wish to test against both positive and negative serial correlation. In that case, there are two possible ways to compute a P value corresponding to a two-tailed test. The first is to assume that the distribution of $\hat{\rho}$ is symmetric, in which case we can use the bootstrap P value

$$\hat{p}^*(\hat{\rho}) = \frac{1}{B} \sum_{j=1}^{B} \text{I}\big(|\rho_j^*| > |\hat{\rho}|\big). \tag{4}$$

This is implicitly a symmetric two-tailed test, since we reject when the fraction of the ρ_j^* that exceed $\hat{\rho}$ in absolute value is small. Alternatively, if we do not assume symmetry, we can use

$$\hat{p}^*(\hat{\rho}) = 2 \min\left(\frac{1}{B} \sum_{j=1}^{B} \text{I}(\rho_j^* \leq \hat{\rho}),\ \frac{1}{B} \sum_{j=1}^{B} \text{I}(\rho_j^* > \hat{\rho}) \right). \tag{5}$$

In this case, for level α, we reject whenever $\hat{\rho}$ is either below the $\alpha/2$ quantile or above the $1 - \alpha/2$ quantile of the empirical distribution of the ρ_j^*. Although tests based on these two P values are both exact, they may yield conflicting results, and their power against various alternatives will differ.

Many common test statistics for serial correlation, heteroskedasticity, skewness, and excess kurtosis in the classical normal linear regression model (2) are pivotal, since they depend on the regressand only through the least squares residuals $\hat{\boldsymbol{u}}$ in a way that is invariant to the scale factor σ. The Durbin-Watson d statistic is a particularly well-known example. We can perform a Monte Carlo test based on d just as easily as a Monte Carlo test based on $\hat{\rho}$, and the two tests should give very similar results. Since we condition on $\boldsymbol{X}$, the infamous upper and lower bounds from the classic tables of the d statistic are quite unnecessary.

With modern computers and appropriate software, it is extremely easy to perform a variety of exact tests in the context of the classical normal linear regression model. These procedures also work when the error terms follow a nonnormal distribution that is known up to a scale factor; we just have to use the appropriate distribution in step 1 above. For further references and a detailed treatment of Monte Carlo tests for heteroskedasticity, see Dufour, Khalaf, Bernard, and Genest (2004).

3. The Parametric Bootstrap

When a hypothesis is tested using a statistic that is not pivotal for that hypothesis, the bootstrap procedure described in the previous section does not lead to an exact Monte Carlo test. However, with a suitable choice of bootstrap DGP, inference that is often more accurate than that provided by conventional asymptotic tests can be performed by use of bootstrap P values. In this section, we focus on the **parametric bootstrap**, in which the bootstrap DGP is completely specified by one or more parameters, some of which have to be estimated.

Examples of the parametric bootstrap are encountered frequently when models are estimated by maximum likelihood. For fixed parameter values, a likelihood function is a probability density which fully characterizes a DGP. The aim of all bootstrap tests is to estimate the distribution of a test statistic under the DGP that generated it, provided that the DGP satisfies the null hypothesis. When a statistic is not pivotal, it is no longer a matter of indifference what DGP is used to generate simulated statistics. Instead, it is desirable to get as good an estimate as possible of the true DGP for the bootstrap DGP. In the context of the parametric bootstrap, this means that we want to estimate the unknown parameters of the true DGP as accurately as possible, since those estimates are used to define the bootstrap DGP.

Consider as an example the probit model, a binary choice model for which each observation y_t on the dependent variable is either 0 or 1. For this model,

$$\Pr(y_t = 1) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta}), \quad t = 1, \ldots, n, \tag{6}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, and $\boldsymbol{X}_t$ is a $1 \times k$ vector of explanatory variables, again treated as fixed. The parameter vector $\boldsymbol{\beta}$ is usually estimated by maximum likelihood (ML).

Suppose that $\boldsymbol{\beta}$ is partitioned into two subvectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and that we wish to test the hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$. The first step in computing a parametric bootstrap P value is to estimate a restricted probit model in which $\boldsymbol{\beta}_2 = \mathbf{0}$, again by ML, so as to obtain restricted estimates $\tilde{\boldsymbol{\beta}}_1$. Next, we compute a suitable test statistic, which would usually be a Lagrange multiplier statistic, a likelihood ratio statistic, or a Wald statistic, although there are other possible choices.

The bootstrap DGP is defined by (6) with $\boldsymbol{\beta}_1 = \tilde{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2 = \mathbf{0}$. Bootstrap samples can be generated easily as follows.

1. Compute the vector of values $\boldsymbol{X}_{t1}\tilde{\boldsymbol{\beta}}_1$, where $\boldsymbol{X}_{t1}$ is the subvector of $\boldsymbol{X}_t$ that corresponds to the nonzero parameters $\boldsymbol{\beta}_1$.
2. For each bootstrap sample, generate a vector of random numbers u_t^*, $t = 1, \ldots, n$, drawn from the standard normal distribution.
3. Set the simulated y_t^* equal to 1 if $\boldsymbol{X}_{t1}\tilde{\boldsymbol{\beta}}_1 + u_t^* > 0$, and to 0 otherwise. By construction, (6) is satisfied for the y_t^*.

For each bootstrap sample, a bootstrap statistic is then computed using exactly the same procedure as the one used to compute the test statistic with the real data. The bootstrap P value is the proportion of bootstrap statistics that are more extreme than the one from the real data. If $\boldsymbol{\beta}_2$ has only one component, the test could be performed using an asymptotic t statistic, and then it would be possible, as we saw in the previous section, to perform one-tailed tests.

Bootstrap DGPs similar to the one described above can be constructed whenever a set of parameter estimates is sufficient to characterize a DGP completely. This is the case for a wide variety of limited dependent models, including more complicated discrete choice models, count data models, and models with censoring, truncation, or sample selection. Hypotheses that can be tested are not restricted to hypotheses about the model parameters. Various specification tests, such as tests for heteroskedasticity or information matrix tests, can be carried out in just the same way.

Why should we expect that a parametric bootstrap test will lead to inference that is more reliable than conventional asymptotic inference? In fact, it does so only if the test statistic τ on which the test is based is **asymptotically pivotal**, which means that its asymptotic distribution, as the sample size $n \to \infty$, is the same for all DGPs that satisfy the null hypothesis. This is not a very strong condition. All statistics that are asymptotically standard normal, or asymptotically chi-squared with known degrees of freedom, or even distributed asymptotically as a Dickey-Fuller distribution, are asymptotically pivotal, since these asymptotic distributions depend on nothing that is specific to a particular DGP in the null hypothesis.

Let us define what in Davidson and MacKinnon (1999a) is called the **rejection probability function**, or **RPF**. The value of this function is the probability that a true null

hypothesis is rejected by a test based on the asymptotic distribution of the statistic τ, as a function of the nominal level α and the parameter vector $\boldsymbol{\theta}$ that characterizes the true DGP. Thus the RPF $R(\alpha, \boldsymbol{\theta})$ satisfies the relation

$$R(\alpha, \boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}\big(\tau \in \text{Rej}(\alpha)\big), \tag{7}$$

where $\text{Rej}(\alpha)$ is the rejection region for an asymptotic test at level α. We use this notation so as to be able to handle one-tailed and two-tailed tests simultaneously. The notation "$\Pr_{\boldsymbol{\theta}}$" indicates that we are evaluating the probability under the DGP characterized by $\boldsymbol{\theta}$. If τ were pivotal, R would not depend on $\boldsymbol{\theta}$.

Suppose now that we carry out a parametric bootstrap test using a realized test statistic $\hat{\tau}$ and a bootstrap DGP characterized by a parameter vector $\boldsymbol{\theta}^*$ that satisfies the null hypothesis. Let $\hat{p}$ be the *asymptotic* P value associated with $\hat{\tau}$; $\hat{p}$ is defined so that $\hat{\tau}$ is on the boundary of the rejection region $\text{Rej}(\hat{p})$. If we let the number B of bootstrap samples tend to infinity, then the bootstrap P value is $R(\hat{p}, \boldsymbol{\theta}^*)$. To see this, observe that this quantity is the probability, according to the bootstrap DGP associated with $\boldsymbol{\theta}^*$, that $\tau \in \text{Rej}(\hat{p})$. For large B, the proportion of bootstrap statistics more extreme than $\hat{\tau}$ tends to this probability.

By a first-order Taylor series approximation around the true parameter vector $\boldsymbol{\theta}_0$,

$$R(\hat{p}, \boldsymbol{\theta}^*) - R(\hat{p}, \boldsymbol{\theta}_0) \cong \boldsymbol{R}^\top(\hat{p}, \boldsymbol{\theta}_0)(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0), \tag{8}$$

where $\boldsymbol{R}(\hat{p}, \boldsymbol{\theta})$ denotes a vector of derivatives of $R(\hat{p}, \boldsymbol{\theta})$ with respect to the elements of $\boldsymbol{\theta}$. The quantity $R(\hat{p}, \boldsymbol{\theta}_0)$ can be interpreted as the ideal P value that we would like to compute if it were possible to do so. Under the DGP associated with $\boldsymbol{\theta}_0$, the probability that $R(\hat{p}, \boldsymbol{\theta}_0)$ is less than α is exactly α. To see this, notice from (7) that

$$R(\alpha, \boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}\big(\tau \in \text{Rej}(\alpha)\big) = \Pr_{\boldsymbol{\theta}}(p < \alpha),$$

where p is the asymptotic P value associated with τ. Thus, under the DGP characterized by $\boldsymbol{\theta}_0$, $R(\alpha, \boldsymbol{\theta}_0)$ is the CDF of p evaluated at α. The random variable $R(p, \boldsymbol{\theta}_0)$, of which $R(\hat{p}, \boldsymbol{\theta}_0)$ is a realization, is therefore distributed as $\text{U}(0, 1)$.

Equation (8) tells us that the difference between the parametric bootstrap P value and the ideal P value is given approximately by the expression $\boldsymbol{R}^\top(\hat{p}, \boldsymbol{\theta}_0)(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)$. Since τ is asymptotically pivotal, the limit of the function $R(\alpha, \boldsymbol{\theta})$ as $n \to \infty$ does not depend on $\boldsymbol{\theta}$. Thus the derivatives in the vector $\boldsymbol{R}(\alpha, \boldsymbol{\theta})$ tend to zero as $n \to \infty$, in regular cases with the same rate of convergence as that of $R(\alpha, \boldsymbol{\theta})$ to its limiting value. This latter rate of convergence is easily seen to be the rate at which the ERP of the asymptotic test tends to zero. The parameter estimates in the vector $\boldsymbol{\theta}^*$ are root-n consistent whenever they are obtained by ML or by any ordinary estimation method. Thus $\boldsymbol{R}^\top(\hat{p}, \boldsymbol{\theta}_0)(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)$, the expectation of which is approximately the ERP of the bootstrap test, tends to zero faster than the ERP of the asymptotic test by a factor of $n^{-1/2}$.

This heuristic argument provides some intuition as to why the parametric bootstrap, when used in conjunction with an asymptotically pivotal statistic, generally yields more reliable inferences than an asymptotic test based on the same statistic. See Beran (1988) for a more rigorous treatment. Whenever the ERP of a bootstrap test declines more rapidly as n increases than that of the asymptotic test on which it is based, the bootstrap test is said to offer **higher-order accuracy** than the asymptotic one, or to benefit from **asymptotic refinements**.

It is important to note that the left-hand side of (8), which is a random variable through the asymptotic P value $\hat{p}$, is not the ERP of the bootstrap test. The bootstrap ERP is rather the expectation of that random variable, which can be seen to depend on the joint distribution of the statistic τ and the estimates $\boldsymbol{\theta}^*$. In some cases, this expectation converges to zero as $n \to \infty$ at a rate even faster than the left-hand side of (8); see Davidson and MacKinnon (1999a) and Davidson and MacKinnon (2004) for more details.

Although the result just given is a very powerful one, it does not imply that a parametric bootstrap test will always be more accurate than the asymptotic test on which it is based. In some cases, an asymptotic test may just happen to perform extremely well even when n is quite small, and the corresponding bootstrap test may perform a little less well. In other cases, neither test may perform at all well in small samples, and the sample size may have to be quite large before the bootstrap test establishes its superiority.

4. Bootstrap DGPs Based on Resampling

The bootstrap, when it was first proposed by Efron (1979, 1982), was an entirely nonparametric procedure. The idea was to draw bootstrap samples from the empirical distribution function (EDF) of the data, a procedure that Efron called **resampling**. Since the EDF assigns probability $1/n$ to each point in the sample, this procedure amounts to drawing each observation of a bootstrap sample randomly, with replacement, from the original sample. Each bootstrap sample thus contains some of the original data points once, some of them more than once, and some of them not at all. Resampling evidently requires the assumption that the data are IID.

In regression models, it is unusual to suppose that the observations are IID. On the other hand, we often suppose that the error terms of a regression model are IID. We do not observe error terms, and so cannot resample them, but we can resample residuals, which are then interpreted as estimates of the error terms. As an example of a bootstrap DGP based on resampling, consider the dynamic linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \gamma y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{9}$$

in which we suppose that y_0 is observed, so that the regression can be run for observations 1 through n. Statistics used to test hypotheses about the parameters $\boldsymbol{\beta}$ and γ are not pivotal for this model. If we assume that the errors are normal, we

can use a parametric bootstrap, as described below, but for the moment we are not willing to make that assumption.

We begin by estimating (9), subject to the restrictions we wish to test, by ordinary or nonlinear least squares, according to the nature of the restrictions. This gives us restricted estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\gamma}$ and a vector of residuals, say $\tilde{\boldsymbol{u}}$. If there is a constant or its equivalent in the regression, then the mean of the elements of $\tilde{\boldsymbol{u}}$ is zero. If not, then it is necessary to center these elements by subtracting their mean, since one of the key assumptions of any regression model is that the errors have an expectation of 0. If we were to resample a set of uncentered residuals, the bootstrap DGP would not belong to the null hypothesis and would give erroneous results.

After recentering, if needed, we can set up a bootstrap DGP as follows. Bootstrap samples are generated recursively from the equation

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + \tilde{\gamma} y_{t-1}^* + u_t^*, \tag{10}$$

where $y_0^* = y_0$, and the u_t^* are resampled from the vector $\tilde{\boldsymbol{u}}$ centered. This is an example of a **semiparametric bootstrap DGP**. The error terms are obtained by resampling, but equation (10) also depends on the parameter estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\gamma}$.

A parametric bootstrap DGP would look just like (10) as regards its recursive nature, but the u_t^* would be generated from the $\mathrm{N}(0, s^2)$ distribution, where s^2 is the least squares variance estimate from the restricted version of (9), that is, $n - k$ times the sum of squared residuals, where k is the number of regression parameters to be estimated under the null hypothesis.

The variance of the resampled bootstrap error terms is the sum of the squared (centered) residuals divided by the sample size n. But this is *not* the least squares estimate s^2, for which one divides by $n - k$. Unless the statistic being bootstrapped is scale invariant, we can get a bootstrap DGP that is a better estimate of the true DGP by rescaling the residuals so as to make their variance equal to s^2. The simplest rescaled residuals are the elements of the vector

$$\dot{\boldsymbol{u}} \equiv \left(\frac{n}{n-k}\right)^{1/2} \tilde{\boldsymbol{u}}, \tag{11}$$

which do indeed have variance s^2. A more sophisticated rescaling method, which takes into account the leverage of each observation, uses the vector with typical element

$$\ddot{u}_t = \lambda \left(\frac{\tilde{u}_t}{(1-h_t)^{1/2}} - \frac{1}{n} \sum_{s=1}^{n} \frac{\tilde{u}_s}{(1-h_s)^{1/2}} \right), \tag{12}$$

where h_t denotes the t^{th} diagonal element of the matrix that projects orthogonally on to the subspace spanned by the regressors of the model used to obtain the residual vector $\tilde{\boldsymbol{u}}$. The second term inside the large parentheses is there to ensure that the rescaled residuals have mean zero, and the factor λ is chosen so that the sample variance of the $\ddot{u}_t$ is equal to s^2. In our experience, methods based on (11) and (12)

generally yield very similar results. However, it may be worth using (12) when a few observations have very high leverage.

The recursive relation (10) that defines the bootstrap DGP must be initialized, and above we chose to do so with the observed value of y_0. This is usually a good choice, and in some cases it is the only reasonable choice. In other cases, the process y_t defined by (10) may have a stationary distribution, and we may be prepared to assume that the observed series y_t is drawn from that stationary distribution. If so, it may be preferable to take for y_0^* a drawing from an estimate of that stationary distribution.

5. Heteroskedasticity

Resampling residuals is reasonable if the error terms are homoskedastic or nearly so. If instead they are heteroskedastic, bootstrap DGPs based on resampled residuals are generally not valid. In this section, we discuss three other methods of constructing bootstrap DGPs, all based on some sort of resampling, which can be used when the error terms of a regression model are heteroskedastic.

To begin with, consider the static linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{13}$$

where $\boldsymbol{\Omega}$ is an unknown, diagonal $n \times n$ covariance matrix. For any model with heteroskedasticity of unknown form, the test statistics which are bootstrapped should always be computed using a heteroskedasticity-consistent covariance matrix estimate, or HCCME. The best-known of these is the one proposed by White (1980) for the model (13), namely,

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{14}$$

where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with squared residuals, possibly rescaled, on the principal diagonal.

The first type of bootstrap DGP that we will discuss is the so-called **pairs bootstrap**, which was originally proposed by Freedman (1981). This is a fully nonparametric procedure that is applicable to a wide variety of models. Unlike resampling residuals, the pairs bootstrap is not limited to regression models. The idea is to resample entire observations from the original data in the form of $[y_t, \boldsymbol{X}_t]$ pairs. Each bootstrap sample consists of some of the original pairs once, some of them more than once, and some of them not at all. This procedure does not condition on $\boldsymbol{X}$ and does not assume that the error terms are IID. Instead, it assumes that all the data are IID drawings from a multivariate distribution, which may permit heteroskedasticity of the y_t conditional on $\boldsymbol{X}_t$.

Although it does not appear to be directly applicable to a dynamic model like (9), the pairs bootstrap can be used with dynamic models that have serially independent

error terms. The idea is simply to treat lagged values of the dependent variable in the same way as other regressors when $\boldsymbol{X}_t$ includes lagged values of y_t. See Gonçalves and Kilian (2004) and the discussion of the block-of-blocks bootstrap in Section 8.

When we use a semiparametric bootstrap DGP like (10), it is generally easy to ensure that the parameter estimates used to define that DGP satisfy the requirements of the null hypothesis. However, when we use the pairs bootstrap, we cannot impose any restrictions on $\boldsymbol{\beta}$. Therefore, we may have to modify the null hypothesis when we calculate the bootstrap test statistics. If the actual null hypothesis is that $\boldsymbol{\beta}_2 = 0$, where $\boldsymbol{\beta}_2$ is a subvector of the regression parameters, we must instead calculate bootstrap test statistics for the null hypothesis that $\boldsymbol{\beta}_2 = \hat{\boldsymbol{\beta}}_2$, where $\hat{\boldsymbol{\beta}}_2$ is the unrestricted estimate. For some specification tests, such as tests for serial correlation, the null imposes no restrictions on $\boldsymbol{\beta}$, and, in such cases, this device is unnecessary.

The trick of changing the null hypothesis so that the bootstrap data automatically satisfy it can be used whenever it is not possible to impose the null hypothesis on the bootstrap DGP. It is also widely used in constructing bootstrap confidence intervals, as we will see in Section 9. However, we recommend imposing the null hypothesis whenever possible, because it generally results in better finite-sample performance. The improvement occurs because restricted estimates are more efficient than unrestricted ones. This improvement is often modest, but it can be substantial in some cases; see Davidson and MacKinnon (1999a).

Flachaire (1999) proposed an alternative version of the pairs bootstrap which makes it possible to impose parametric restrictions. First, the regression model is estimated incorporating the restrictions of the null hypothesis. This yields restricted estimates $\tilde{\boldsymbol{\beta}}$ and restricted residuals $\tilde{u}_t$. The bootstrap DGP then resamples the pairs $[\tilde{u}_t, \boldsymbol{X}_t]$ and reconstructs the bootstrap dependent variable y_t^* by the formula

$$y_t^* = \boldsymbol{X}_t^* \tilde{\boldsymbol{\beta}} + u_t^*,$$

where each pair $[u_t^*, \boldsymbol{X}_t^*]$ is randomly resampled from the set of pairs $[\tilde{u}_t, \boldsymbol{X}_t]$. This bootstrap DGP accounts for heteroskedasticity conditional on the regressors, imposes the parametric restrictions of the null hypothesis, and uses restricted residuals, which presumably provide better estimates of the unobserved error terms. Flachaire gives some limited simulation results in which tests based on his modified pairs bootstrap have a smaller ERP than ones based on the conventional pairs bootstrap.

The pairs bootstrap is not the only type of bootstrap DGP that allows for heteroskedasticity of unknown form in a regression model. A very different technique called the **wild bootstrap** is also available. It very often seems to work better than the pairs bootstrap when it is applicable; see MacKinnon (2002) and Flachaire (2004) for some simulation evidence. The wild bootstrap is a semiparametric procedure, in some ways quite similar to the one discussed in Section 4, but the IID assumption is not imposed by the method used to simulate the bootstrap errors. Key early references are Wu (1986), Liu (1988), and Mammen (1993).

For testing restrictions on the model (13), the wild bootstrap DGP would be

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + f(\tilde{u}_t)v_t^*, \tag{15}$$

where $\tilde{\boldsymbol{\beta}}$ denotes the least-squares estimates subject to the restrictions being tested, $f(\tilde{u}_t)$ is a transformation of the t^{th} restricted residual $\tilde{u}_t$, and v_t^* is a random variable with mean 0 and variance 1. The simplest choice for $f(\cdot)$ is just $f(\tilde{u}_t) = \tilde{u}_t$, but another natural choice is

$$f(\tilde{u}_t) = \frac{\tilde{u}_t}{(1-h_t)^{1/2}},$$

which ensures that the $f(\hat{u}_t)$ would have constant variance if the error terms were homoskedastic.

There are, in principle, many ways to specify v_t^*. The most popular approach is to use the two-point distribution

$$F_1: \quad v_t^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with prob. } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with prob. } (\sqrt{5}-1)/(2\sqrt{5}), \end{cases}$$

which was suggested by Mammen (1993). However, a much simpler two-point distribution is the **Rademacher distribution**

$$F_2: \quad v_t^* = \begin{cases} -1 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Davidson and Flachaire (2001) have shown, on the basis of both theoretical analysis and simulation experiments, that wild bootstrap tests based on F_2 usually perform better than ones based on F_1, especially when the conditional distribution of the error terms is approximately symmetric.

The error terms for the wild bootstrap DGP (15) do not look very much like those for the true DGP (13). When a two-point distribution is used, the bootstrap error term can take on only two possible values for each observation. With F_2, these are just plus and minus $f(\tilde{u}_t)$. Nevertheless, the wild bootstrap apparently mimics the essential features of many actual DGPs well enough for it to be useful in many cases.

It is possible to use the wild bootstrap with dynamic models. What Gonçalves and Kilian (2004) call the **recursive-design wild bootstrap** is simply a bootstrap DGP that combines recursive calculation of the regression function, as in (10), with wild bootstrap error terms, as in (15). They provide theoretical results to justify the use of the F_1 form of the recursive-design wild bootstrap for pure autoregressive models with ARCH errors, and they provide simulation evidence that symmetric confidence intervals work well for AR(1) models with ARCH errors.

In a related paper, Godfrey and Tremayne (2004) have provided evidence that heteroskedasticity-robust tests for serial correlation in dynamic linear regression models like (9) perform markedly better when they are bootstrapped using either the F_1

or F_2 form of the recursive-design wild bootstrap than when asymptotic critical values are used.

Although the wild bootstrap often works well, it is possible to find combinations of $\boldsymbol{X}$ matrix and pattern of heteroskedasticity for which tests and/or confidence intervals based on it are not particularly reliable in finite samples; see, for example, MacKinnon (2002). Problems are most likely to arise when there is severe heteroskedasticity and a few observations have exceptionally high leverage.

The pairs bootstrap and the wild bootstrap are primarily used when it is thought that the heteroskedasticity is conditional on the regressors. For conditional heteroskedasticity of the ARCH/GARCH variety, a parametric or semiparametric bootstrap DGP can be used instead. The GARCH(p, q) process is defined by $p+q+1$ parameters, of which one is a scale factor. All these parameters can be estimated consistently. As an example, consider the GARCH(1,1) process, defined by the recurrence

$$u_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0,1), \quad \sigma_t^2 = \alpha + \gamma u_{t-1}^2 + \delta \sigma_{t-1}^2. \tag{16}$$

For a bootstrap DGP, we may use estimates of the GARCH parameters α, β, and γ, and, for the ε_t^*, we may use either independent standard normal random numbers, if we prefer a parametric bootstrap, or resampled residuals, recentered if necessary, and rescaled so as to have variance 1.

As with all recursive relationships, (16) must be initialized. In fact, we need values for both u_1^2 and σ_1^2 in order to use it to generate a GARCH(1,1) process. If u_1 is observed, or if a residual $\hat{u}_1$ can be computed, then it makes sense to use it to initialize (16). For σ_1^2, a good choice in most circumstances is to use an estimate of the stationary variance of the process, here $\alpha/(1-\gamma-\delta)$.

6. Covariance Matrices and Bias Correction

If we generate a number of bootstrap samples and use each of them to estimate a parameter vector, it seems natural to use the sample covariance matrix of the bootstrap parameter estimates as an estimator of the covariance matrix of the original parameter estimates. In fact, early work such as Efron (1982) emphasized the use of the bootstrap primarily for this purpose.

Although there are cases in which **bootstrap covariance matrices**, or **bootstrap standard errors**, are useful, that is not true for regression models. Consider the semiparametric bootstrap DGP (10) that we discussed in Section 4, in which the bootstrap error terms are obtained by resampling rescaled residuals. Suppose we use this bootstrap DGP with the static linear regression model (13). If $\bar{\boldsymbol{\beta}}^*$ denotes the sample mean of the bootstrap parameter estimates $\hat{\boldsymbol{\beta}}_j^*$, then the sample covariance matrix of the $\hat{\boldsymbol{\beta}}_j^*$ is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_j^*) = \frac{1}{B} \sum_{j=1}^{B} (\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)^\top. \tag{17}$$

The probability limit of this **bootstrap covariance matrix**, as $B \to \infty$, is

$$\operatorname*{plim}_{B\to\infty}\left(\frac{1}{B}\sum_{j=1}^{B}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}_j^*\boldsymbol{u}_j^{*\top}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right) = \sigma_*^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{18}$$

where σ_*^2 is the variance of the bootstrap errors, which should be equal to s^2 if the error terms have been rescaled properly before resampling.

This example makes two things clear. First, it is just as necessary to make an appropriate choice of bootstrap DGP for covariance matrix estimation as for hypothesis testing. In the presence of heteroskedasticity, the semiparametric bootstrap DGP (10) is not appropriate, because (18) is not a valid estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$. Second, when the errors are homoskedastic, so that (18) is valid, it is neither necessary nor desirable to use the bootstrap to calculate a covariance matrix. Every OLS regression package can calculate the matrix $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. The semiparametric bootstrap simply replaces s^2 by an estimate that converges to it as $B \to \infty$.

The pairs bootstrap does lead to a valid covariance matrix estimator for the model (13). In fact, it can readily be shown that, as $B \to \infty$, the bootstrap estimate (17) tends to the White estimator (14); see Flachaire (2002) for details. It therefore makes little sense to use the pairs bootstrap when it is so easy to calculate the HCCME (14) without doing any simulation at all. In general, it makes sense to calculate covariance matrices via the bootstrap only when it is very difficult to calculate reliable ones in any other way. The linear regression model is not such a case.

This raises an important theoretical point. Even when it does make sense to compute a bootstrap standard error, a test based on it does not benefit from the asymptotic refinements that accrue to a bootstrap test based on an asymptotically pivotal test statistic. Consider a test statistic of the form

$$\frac{\hat{\theta} - \theta_0}{s_\theta^*}, \tag{19}$$

where $\hat{\theta}$ is a parameter estimate, θ_0 is the true value, and s_θ^* is a bootstrap standard error. When $n^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically normal, and the bootstrap DGP yields a valid standard error estimate, the statistic (19) is asymptotically distributed as $\mathrm{N}(0,1)$. However, there is, in general, no reason to suppose that it yields more accurate inferences in finite samples than a similar statistic that uses some other standard error estimate.

It might seem natural to modify (19) by using a bias-corrected estimate of θ instead of $\hat{\theta}$. Suppose that $\bar{\theta}^*$ is the mean of a set of bootstrap estimates θ_j^* obtained by using a parametric or semiparametric bootstrap DGP characterized by the parameter $\hat{\theta}$. Then a natural estimate of bias is just $\bar{\theta}^* - \hat{\theta}$. This implies that a **bias-corrected** estimate is

$$\hat{\theta}^* \equiv \hat{\theta} - (\bar{\theta}^* - \hat{\theta}) = 2\hat{\theta} - \bar{\theta}^*. \tag{20}$$

In most cases, $\hat{\theta}^*$ is less biased than $\hat{\theta}$. However, the variance of $\hat{\theta}^*$ is

$$\mathrm{Var}(\hat{\theta}^*) = 4\,\mathrm{Var}(\hat{\theta}) + \mathrm{Var}(\bar{\theta}^*) - 4\,\mathrm{Cov}(\hat{\theta}, \hat{\theta}^*),$$

which is greater than $\mathrm{Var}(\hat{\theta})$ except in the extreme case in which

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}^*) = \mathrm{Cov}(\hat{\theta}, \hat{\theta}^*).$$

In general, using the bootstrap to correct bias results in increased variance. MacKinnon and Smith (1998) propose some alternative methods of simulation-based bias correction, which sometimes work better than (20). Davison and Hinkley (1997) discuss a number of other bias-correction methods.

7. More than one Dependent Variable

Although there is little difficulty in adapting the methods described so far to systems of equations that define more than one dependent variable, it is not so simple to set up adequate bootstrap DGPs when we are interested in only one equation of such a system. Econometricians very frequently estimate a regression model using instrumental variables in order to take account of possible endogeneity of the explanatory variables. If some or all of the explanatory variables are indeed endogenous, then they, as well as the dependent variable of the regression being estimated, must be explicitly generated by a bootstrap DGP. The question that arises is just how this should be done.

Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t \equiv \boldsymbol{Z}_t\boldsymbol{\beta} + \boldsymbol{Y}_t\boldsymbol{\gamma} + u_t, \tag{21}$$

where the variables in $\boldsymbol{Z}_t$ are treated as exogenous and those in $\boldsymbol{Y}_t$ as endogenous, that is, correlated with the error term u_t. If we form a set of instrumental variables $\boldsymbol{W}_t$ where $\boldsymbol{W}_t$ contains $\boldsymbol{Z}_t$ as a subvector and has at least as many additional exogenous elements as there are endogenous variables in $\boldsymbol{Y}_t$, then the IV estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} \equiv (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y} \tag{22}$$

is root-n consistent under standard regularity conditions, with asymptotic covariance matrix

$$\mathrm{Var}\big(\operatorname*{plim}_{n\to\infty} n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}} - \boldsymbol{\beta}_0)\big) = \sigma_0^2 \operatorname*{plim}_{n\to\infty} (n^{-1}\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1},$$

where $\boldsymbol{\beta}_0$ is the true parameter vector and σ_0^2 the true error variance. The estimator (22) is asymptotically efficient in the class of IV estimators if the endogenous variables $\boldsymbol{Y}_t$ are related to the instruments by the set of linear relations

$$\boldsymbol{Y}_t = \boldsymbol{W}_t\boldsymbol{\Pi} + \boldsymbol{V}_t, \tag{23}$$

where the error terms $\boldsymbol{V}_t$ are of mean zero and are, in general, correlated with u_t.

If we are willing to assume that both (21) and (23) are correctly specified, then we can treat them jointly as a system of equations simultaneously determining y_t and $\boldsymbol{Y}_t$. The parameters of the system can all be estimated, $\boldsymbol{\beta}$ by either (22) or a restricted version of it if we wish to test a set of restrictions, and $\boldsymbol{\Pi}$ by least squares applied to the reduced-form equations (23). The covariance matrix of u_t and $\boldsymbol{V}_t$, under the assumption that the pairs $[u_t, \boldsymbol{V}_t]$ are IID, can be estimated using the squares and cross-products of the residuals given by estimating (21) and (23).

If we let the estimated covariance matrix be $\hat{\boldsymbol{\Sigma}}$, then a possible parametric bootstrap DGP is

$$y_t^* = \boldsymbol{Z}_t\hat{\boldsymbol{\beta}} + \boldsymbol{Y}_t^*\hat{\boldsymbol{\gamma}} + u_t^*, \quad \boldsymbol{Y}_t^* = \boldsymbol{W}_t\hat{\boldsymbol{\Pi}} + \boldsymbol{V}_t^*, \quad \begin{bmatrix} u_t^* \\ \boldsymbol{V}_t^* \end{bmatrix} \sim \text{NID}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}). \tag{24}$$

Similarly, a possible semiparametric bootstrap DGP looks just like (24) except that the bootstrap errors $[u_t^*, \boldsymbol{V}_t^*]$ are obtained by resampling from the pairs $[\hat{u}_t, \hat{\boldsymbol{V}}_t]$ of residuals from the estimation of (21) and (23). If there is no constant in the set of instruments, then it is necessary to recenter these residuals before resampling. Some sort of rescaling procedure could also be used, but it is not at all clear whether doing so would have any beneficial effect.

Another nonparametric approach to bootstrapping a model like (21) is to extend the idea of the pairs bootstrap and construct bootstrap samples by resampling from the tuples $[y_t, \boldsymbol{X}_t, \boldsymbol{W}_t]$. This method makes very weak assumptions about the joint distribution of these variables, but it does assume that they are IID across observations. As we saw in Section 5, the IID assumption allows for heteroskedasticity conditional on the exogenous variables $\boldsymbol{W}_t$. It should be possible to incorporate restrictions just as with Flachaire's modification of the pairs bootstrap, although, to the best of our knowledge, this has not yet been done.

The parametric and semiparametric bootstrap procedures described above for the model specified by equations (21) and (23) can easily be extended to deal with any fully specified set of seemingly unrelated equations or simultaneous equation system. As long as the model provides a mechanism for generating *all* of its endogenous variables, and the parameters can be consistently estimated, a bootstrap DGP for such a model is conceptually no harder to set up than for a single-equation model. However, not much is yet known about the finite-sample properties of bootstrap procedures in multivariate models. See Rilstone and Veall (1996), Inoue and Kilian (2002), and MacKinnon (2002) for limited evidence about a few particular cases.

8. Bootstrap DGPs for Dependent Data

The bootstrap DGPs that we have discussed so far are not valid when applied to models with dependent errors having an unknown pattern of dependence. For such models, we wish to specify a bootstrap DGP which generates correlated error terms that exhibit approximately the same pattern of dependence as the real errors, even though we do not know the process that actually generated the errors. There are two main approaches, neither of which is entirely satisfactory in all cases.

The first approach is a semiparametric one called the **sieve bootstrap**. It is based on the fact that any linear, invertible time-series process can be approximated by an AR(∞) process. The idea is to estimate a stationary AR(p) process and use this estimated process, perhaps together with resampled residuals from the estimation of the AR(p) process, to generate bootstrap samples. For example, suppose we are concerned with the static linear regression model (13), but the covariance matrix $\boldsymbol{\Omega}$ is no longer assumed to be diagonal. Instead, it is assumed that $\boldsymbol{\Omega}$ can be well approximated by the covariance matrix of a stationary AR(p) process, which implies that the diagonal elements are all the same.

In this case, the first step is to estimate the regression model, possibly after imposing restrictions on it, so as to generate a parameter vector $\hat{\boldsymbol{\beta}}$ and a vector of residuals $\hat{\boldsymbol{u}}$ with typical element $\hat{u}_t$. The next step is to estimate the AR(p) model

$$\hat{u}_t = \sum_{i=1}^{p} \rho_i \hat{u}_{t-i} + \varepsilon_t \tag{25}$$

for $t = p+1, \ldots, n$. In theory, the order p of this model should increase at a certain rate as the sample size increases. In practice, p is most likely to be determined either by using an information criterion like the AIC or by sequential testing. Care should be taken to ensure that the estimated model is stationary. This may require the use of full maximum likelihood to estimate (25), rather than least squares.

Estimation of (25) yields residuals and an estimate $\hat{\sigma}_\varepsilon^2$ of the variance of the ε_t, as well as the estimates $\hat{\rho}_i$. We may use these to set up a variety of possible bootstrap DGPs, all of which take the form

$$y_t^* = \boldsymbol{X}_t \hat{\boldsymbol{\beta}} + u_t^*.$$

There are two choices to be made, namely, the choice of parameter estimates $\hat{\boldsymbol{\beta}}$ and the generating process for the bootstrap errors u_t^*. One choice for $\hat{\boldsymbol{\beta}}$ is just the OLS estimates from running (13). But these estimates, although consistent, are not efficient if $\boldsymbol{\Omega}$ is not a scalar matrix. We might therefore prefer to use feasible GLS estimates. An estimate $\hat{\boldsymbol{\Omega}}$ of the covariance matrix can be obtained by solving the Yule-Walker equations, using the $\hat{\rho}_i$ in order to obtain estimates of the auto-covariances of the AR(p) process. Then a Cholesky decomposition of $\hat{\boldsymbol{\Omega}}^{-1}$ provides the feasible GLS transformation to be applied to the dependent variable $\boldsymbol{y}$ and the

explanatory variables $\boldsymbol{X}$ in order to compute feasible GLS estimates of $\boldsymbol{\beta}$, restricted as required by the null hypothesis under test.

For observations after the first p, the bootstrap errors are generated as follows:

$$u_t^* = \sum_{i=1}^{p} \hat{\rho}_i u_{t-i}^* + \varepsilon_t^*, \quad t = p+1, \ldots, n, \tag{26}$$

where the ε_t^* can either be drawn from the $\mathrm{N}(0, \hat{\sigma}_\varepsilon^2)$ distribution for a parametric bootstrap or resampled from the residuals $\hat{\varepsilon}_t$ from the estimation of (25), preferably rescaled by the factor $\sqrt{n/(n-p)}$. Before we can use (26), of course, we must generate the first p bootstrap errors, the u_t^*, for $t = 1, \ldots, p$.

One way to do so is just to set $u_t^* = \hat{u}_t$ for the first p observations of each bootstrap sample. This is analogous to what we proposed for the bootstrap DGP (10) used in conjunction with the dynamic model (9): We initialize (26) with fixed starting values given by the real data. Unless we are sure that the AR(p) process is really stationary, rather than just being characterized by values of the ρ_i that correspond to a stationary covariance matrix, this is the only appropriate procedure.

If we are happy to impose full stationarity on the bootstrap DGP, then we may draw the first p values of the u_t^* from the p–variate stationary distribution. This is easy to do if we have solved the Yule-Walker equations for the first p autocovariances, provided that we assume normality. If normality is an uncomfortably strong assumption, then we can initialize (26) in any way we please and then generate a reasonably large number (say 200) of bootstrap errors recursively, using resampled rescaled values of the $\hat{\varepsilon}_t$ for the ε_t^*. We then throw away all but the last p of these errors and use those to initialize (26). In this way, we approximate a stationary process with the correct estimated stationary covariance matrix, but with no assumption of normality.

The sieve bootstrap method has been used to improve the finite-sample properties of unit root tests by Park (2003) and Chang and Park (2003), but it has not yet been widely used in econometrics. The fact that it does not allow for heteroskedasticity is a limitation. Moreover, AR(p) processes do not provide good approximations to every time-series process that might arise in practice. An example for which the approximation is exceedingly poor is an MA(1) process with a parameter close to -1. The sieve bootstrap cannot be expected to work well in such cases. For more detailed treatments, see Bühlmann (1997, 2002), Choi and Hall (2000), and Park (2002).

The second principal method of dealing with dependent data is the **block bootstrap**, which was originally proposed by Künsch (1989). This method is much more widely used than the sieve bootstrap. The idea is to divide the quantities that are being resampled, which might be either rescaled residuals or $[\boldsymbol{y}, \boldsymbol{X}]$ pairs, into blocks of b consecutive observations, and then resample the blocks. The blocks may be either overlapping or nonoverlapping. In either case, the choice of block length, b, is evidently very important. If b is small, the bootstrap samples cannot possibly mimic the patterns of dependence in the original data, because these patterns are broken

whenever one block ends and the next begins. However, if b is large, the bootstrap samples will tend to be excessively influenced by the random characteristics of the actual sample.

For the block bootstrap to work asymptotically, the block length must increase as the sample size n increases, but at a slower rate, which varies depending on what the bootstrap samples are to be used for. In some common cases, b should be proportional to $n^{1/3}$, but with a factor of proportionality that is, in practice, unknown. Unless the sample size is very large, it is generally impossible to find a value of b for which the bootstrap DGP provides a really good approximation to the unknown true DGP.

A variation of the block bootstrap is the **stationary bootstrap** proposed by Politis and Romano (1994), in which the block length is random rather than fixed. This procedure is commonly used in practice. However, Lahiri (1999) provides both theoretical arguments and limited simulation evidence which suggest that fixed block lengths are better than variable ones and that overlapping blocks are better than nonoverlapping ones. Thus, at the present time, the procedure of choice appears to be the **moving-block bootstrap**, in which there are $n-b+1$ blocks, the first containing observations 1 through b, the second containing observations 2 through $b+1$, and the last containing observations $n-b+1$ through n.

It is possible to use block bootstrap methods with dynamic models. For example, consider the dynamic linear regression model (9). Let

$$\boldsymbol{Z}_t \equiv [y_t, y_{t-1}, \boldsymbol{X}_t].$$

For this model, we could construct $n-b+1$ overlapping blocks

$$\boldsymbol{Z}_1 \dots \boldsymbol{Z}_b, \;\; \boldsymbol{Z}_2 \dots \boldsymbol{Z}_{b+1}, \dots\dots, \boldsymbol{Z}_{n-b+1} \dots \boldsymbol{Z}_n$$

and resample from them. This is the moving-block analog of the pairs bootstrap. When there are no exogenous variables and several lagged values of the dependent variable, the $\boldsymbol{Z}_t$ are themselves blocks of observations. Therefore, this method is sometimes referred to as the **block-of-blocks bootstrap**. Notice that, when the block size is 1, the block-of-blocks bootstrap is simply the pairs bootstrap adapted to dynamic models, as in Gonçalves and Kilian (2004).

Block bootstrap methods are conceptually simple. However, there are many different versions, most of which we have not discussed, and theoretical analysis of their properties tends to require advanced techniques. The biggest problem with block bootstrap methods is that they often do not work very well. We have already provided an intuitive explanation of why this is the case. From a theoretical perspective, the problem is that, even when the block bootstrap offers higher-order accuracy than asymptotic methods, it often does so to only a modest extent. The improvement is always of higher order in the independent case, where blocks should be of length 1, than in the dependent case, where the block size must be greater than 1 and must increase at an optimal rate with the sample size. See Hall, Horowitz, and Jing (1995) and Andrews (2002, 2004), among others.

There are several valuable, recent surveys of bootstrap methods for time-series data. These include Bühlmann (2002), Politis (2003), and Härdle, Horowitz, and Kreiss (2003). Surveys that are older or deal with methods for time-series data in less depth include Li and Maddala (1996), Davison and Hinkley (1997, Chapter 8), Berkowitz and Kilian (2000), Horowitz (2001), and Horowitz (2003).

9. Confidence Intervals

A confidence interval at level $1 - \alpha$ for some parameter θ can be constructed as the set of values of θ_0 such that the hypothesis $\theta = \theta_0$ is not rejected by a test at level α. This suggests that confidence intervals can be constructed using bootstrap methods, and that is indeed the case. Suppose that $\hat{\theta}$ is an estimate of θ, and $\hat{s}_\theta$ is its estimated standard error. Then, in many cases, the asymptotic t statistic

$$\tau = \frac{\hat{\theta} - \theta_0}{\hat{s}_\theta} \tag{27}$$

is pivotal or asymptotically pivotal when θ_0 is the true value of θ. As an example, θ might be one of the regression parameters of a classical normal linear regression model like (2). In this case, if $\hat{\theta}$ is the OLS estimator of θ, and $\hat{s}_\theta$ is the usual OLS standard error, we know that τ follows Student's t distribution with a degrees-of-freedom parameter that depends only on the sample size.

When the distribution of τ is known, we can find the $\alpha/2$ and $1 - \alpha/2$ quantiles of that distribution, say $t_{\alpha/2}$ and $t_{1-\alpha/2}$, and use them to construct the confidence interval

$$\big[\hat{\theta} - \hat{s}_\theta t_{1-\alpha/2},\ \ \hat{\theta} - \hat{s}_\theta t_{\alpha/2}\big]. \tag{28}$$

This interval contains all the values of θ_0 that satisfy the inequalities

$$t_{\alpha/2} \le \frac{\hat{\theta} - \theta_0}{\hat{s}_\theta} \le t_{1-\alpha/2}. \tag{29}$$

If θ_0 is the true parameter value, then the probability that $\hat{\theta}$ satisfies (29) is exactly $1 - \alpha$, by construction, and so the coverage probability of the interval (28) is also $1 - \alpha$. At first glance, the confidence interval (28) may seem odd, because the lower limit depends on an upper-tail quantile and the upper limit depends on a lower-tail quantile. This seeming inversion is not necessary when τ has a symmetric distribution, and is sometimes hidden when (28) is written in other ways which may be more familiar, but it is essential with an asymmetric distribution.

Whether or not the distribution of τ is known, we can replace $t_{\alpha/2}$ and $t_{1-\alpha/2}$ by the corresponding quantiles of the empirical distribution of B bootstrap statistics t_j^*. It is important to note that the bootstrap statistics must test a hypothesis that is true of the bootstrap DGP. This point was discussed in connection with the conventional

(Freedman, 1981) version of the pairs bootstrap, in which the hypothesis tested is that $\theta = \hat{\theta}$, not $\theta = \theta_0$.

If $\frac{1}{2}\alpha(B+1)$ is an integer, and if τ is an exact pivot, using the quantiles of a bootstrap distribution leads to a confidence interval with coverage probability of exactly $1-\alpha$, just as a P value based on an exact pivot gives a rejection probability equal to a desired significance level α if $\alpha(B+1)$ is an integer. In all cases, the bootstrap quantiles are calculated as the order statistics of rank $(\alpha/2)(B+1)$ and $(1-\alpha/2)(B+1)$ in the set of the t_j^* sorted from smallest to largest. If $B = 199$ and $\alpha = .05$, for example, the empirical quantiles are t_5^* and t_{195}^*. Thus a confidence interval comparable to (28) has the form

$$\big[\hat{\theta} - \hat{s}_\theta t^*_{(1-\alpha/2)(B+1)},\ \ \hat{\theta} - \hat{s}_\theta t^*_{\alpha/2(B+1)}\big]. \tag{30}$$

Such an interval is called a **Monte Carlo confidence interval** if τ is an exact pivot, and a **bootstrap confidence interval** otherwise. The cost of using quantiles estimated with a finite number of bootstrap statistics rather than the true quantiles of the distribution of τ is that, on average, confidence intervals constructed using the former are longer than ones constructed using the latter.

If the distribution of τ is believed to be symmetric around the origin, we can use the bootstrap confidence interval

$$\big[\hat{\theta} - \hat{s}_\theta |t^*|_{(1-\alpha)(B+1)},\ \ \hat{\theta} + \hat{s}_\theta |t^*|_{(1-\alpha)(B+1)}\big]. \tag{31}$$

instead of (30). Here $|t^*|_{(1-\alpha)(B+1)}$ denotes number $(1-\alpha)(B+1)$ in the sorted list of the absolute values of the t_j^*. This **symmetric confidence interval** is related to a test based on the bootstrap P value (4) in the same way that the **equal-tailed confidence interval** (30) is related to a test based on (5).

The method of constructing bootstrap confidence intervals described above is often called the **bootstrap *t* method** or **percentile *t* method**, because it involves percentiles (that is, quantiles) of the distribution of bootstrap t statistics. If τ is merely asymptotically pivotal, bootstrap t confidence intervals are subject to coverage error. But just as bootstrap P values based on an asymptotic pivot have an ERP that benefits from asymptotic refinements, so do the coverage errors of bootstrap t confidence intervals decline more rapidly as the sample size increases than those of confidence intervals based on the nominal asymptotic distribution of τ; see Hall (1992).

The confidence interval (28) is obtained by "inverting" the test for which the t statistic (27) is the test statistic, in the sense that the interval contains exactly those values of θ_0 for which a two-tailed test of the hypothesis $\theta = \theta_0$ based on (27) is not rejected at level α. If instead we chose to invert a one-tailed test, we would obtain a confidence interval open to infinity in one direction. In this case, we would base the interval on the α or $1-\alpha$ quantile of the bootstrap distribution.

The inversion of the test based on (27) is particularly easy to carry out because (27) depends linearly on θ_0. This is generally true if one uses an asymptotic t statistic. But

such statistics are associated with Wald tests, and they may therefore suffer from the well-known disadvantages of Wald tests. It is not necessary to limit oneself to Wald tests when constructing confidence intervals. Davison, Hinkley, and Young (2003), for example, discuss the construction of confidence intervals based on inverting the signed square root of a likelihood ratio statistic. In general, let $\tau(\boldsymbol{y}, \theta_0)$ denote a test statistic that depends on data $\boldsymbol{y}$ and tests the hypothesis that $\theta = \theta_0$. For any given distribution of $\tau(\boldsymbol{y}, \theta_0)$ under the null hypothesis, exact, asymptotic, or bootstrap, the (two-tailed) confidence interval obtained by inverting the test based on $\tau(\boldsymbol{y}, \theta_0)$ is the set of values of θ_0 that satisfy the inequalities

$$t_{\alpha/2} \leq \tau(\boldsymbol{y}, \theta_0) \leq t_{1-\alpha/2},$$

where, as before, $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are quantiles of the given distribution. It is clear that, when $\tau(\boldsymbol{y}, \theta)$ is a nonlinear function of θ, solving the equations

$$\tau(\boldsymbol{y}, \theta_+) = t_{\alpha/2} \text{ and } \tau(\boldsymbol{y}, \theta_-) = t_{1-\alpha/2}$$

that implicitly define the upper and lower limits $\theta_\pm$ of the confidence interval may be more computationally demanding than solving the equations that result from (29) in order to obtain the interval (28).

A great many other procedures have been proposed for constructing bootstrap confidence intervals. These include two very different procedures that are both confusingly called the **percentile method** by different authors. Neither of these is to be recommended in most cases, because they both involve inverting quantities that are not even asymptotically pivotal; see Hall (1992). They also include a number of more complicated techniques, such as the **grid bootstrap** of Hansen (1999). References that discuss a variety of methods for constructing confidence intervals include DiCiccio and Efron (1996) and Davison and Hinkley (1997). For reasons of space, however, we will not discuss any of them.

A bootstrap t confidence interval may be unreliable if τ is too far from being pivotal in finite samples. If so, a natural way to obtain a more reliable interval is to invert a test statistic that is closer to being pivotal. An approach that avoids the computational cost of inverting something other than a Wald test is to apply a nonlinear transformation to the parameter of interest, form a confidence interval for the transformed parameter, and then map from that interval to one for the original parameter. This can work well if the t statistic for the transformed parameter is closer to being pivotal than the one for the original parameter.

10. The Performance of Bootstrap Methods

The bootstrap, whether used for hypothesis testing or the construction of confidence intervals, relies on the choice of a suitable bootstrap DGP for generating simulated data. We want the simulated data to have statistical properties as close as possible to those of the actual data, under the assumption that the latter were generated by a

DGP that satisfies the requirements of the hypothesis under test or of the model for the parameters of which confidence intervals are sought. Consequently, we have tried to emphasize the importance of choosing a bootstrap DGP adapted to the problem at hand. Problems can and do arise if it is difficult or impossible to find a suitable bootstrap DGP. However, for many commonly used econometric models, it is not hard to do so if one takes a modest amount of care.

In this chapter, we have largely confined our discussion to linear models. This has been purely in the interests of clarity. Nonlinear regression models, with or without heteroskedasticity or serial correlation, can be handled using the sorts of bootstrap DGPs we have described. The same is true of multivariate nonlinear systems. The only disadvantage is that computing times are longer when nonlinear estimation is involved, and even this disadvantage can be minimized by use of techniques that we describe in Davidson and MacKinnon (1999b).

For a Monte Carlo test based on an exactly pivotal quantity, any DGP belonging to the model for which that quantity is pivotal can serve as the bootstrap DGP. We have seen that Monte Carlo tests are exact, and that Monte Carlo confidence intervals have exact coverage, if B is chosen properly. Intuitively, then, we expect bootstrapping to perform better the closer it is to a Monte Carlo procedure. This means that the quantity that is bootstrapped should be as close as possible to being pivotal, and that the bootstrap DGP should be as good an estimate as possible of the true DGP. As we saw in Section 3, asymptotic refinements are available for the bootstrap when both these requirements are met. This is the case for the parametric bootstrap, which can be used with almost any fully parametric model. It is a natural choice if estimation is by maximum likelihood, but it makes sense only if one is confident of the specification of the model.

Once we get over the hurdle of finding a suitable bootstrap DGP, the delicate part of bootstrapping is over, since we can use the general techniques laid out in this chapter for using bootstrap samples to generate P values or confidence intervals.

In Section 2, we saw that exact Monte Carlo procedures are available for univariate linear regression models with fixed regressors and IID normal errors, but that bootstrap methods which allow for lagged dependent variables and/or nonnormal errors are no longer exact. If we can use a parametric bootstrap, using reasonably precise estimates of the nuisance parameters on which the distribution of the test statistic depends, bootstrap tests and confidence intervals can be remarkably accurate. In fact, numerous simulation experiments suggest that, for univariate regression models with IID errors, bootstrap methods generally work extremely well. In particular, this seems to be true for serial correlation tests (MacKinnon, 2002), tests of common factor restrictions (Davidson and MacKinnon, 1999b), and nonnested hypothesis tests (Godfrey, 1998; Davidson and MacKinnon, 2002). It would be surprising if it were not true for any sort of test on the parameters of a linear or nonlinear regression function, except perhaps in extreme cases like some of the ones considered by Davidson and MacKinnon (2002).

Once we move out of the realm of IID errors, the performance of bootstrap methods becomes harder to predict. The pairs bootstrap is very generally applicable when the data are independent, but its finite-sample performance can leave a lot to be desired; see, for example, MacKinnon (2002). The wild bootstrap is less widely applicable than the pairs bootstrap, but it generally outperforms the latter, especially when the F_2 variant is used. However, it is generally not as reliable as resampling rescaled residuals in the IID case.

With dependent data, bootstrap methods often do not perform well at all. Neither the sieve bootstrap nor the best available block bootstrap methods can be relied upon to yield accurate inferences in samples of moderate size. Even for quite large samples, they may perform little better than asymptotic tests, although there are cases in which they do perform well. At this stage, all we can recommend is that practitioners should, if possible, conduct their own simulation experiments, for the specific model and test(s) they are interested in, to see directly whether the available bootstrap procedures seem to yield reliable inferences.

Much modern bootstrap research deals with **bootstrap failure**, by which we mean that a bootstrap DGP gives such a poor approximation to the true DGP that bootstrap inference is severely misleading. It should be noted that a failure of one type of bootstrap DGP does not imply that all bootstrap methods are bound to fail; in many cases, a bootstrap failure has led to the development of more powerful methods. One case in which bootstrap failure can be a serious problem in applied work is when the true DGP generates random variables with fat tails. For instance, as long ago as 1987, Athreya (1987) showed that resampling from data generated by a distribution with an infinite variance does not allow asymptotically valid inference about the mean of that distribution. Although better methods have been developed since then, fat tails still constitute a serious challenge for conventional bootstrap techniques.

There is an enormous variety of methods for constructing bootstrap DGPs that we have not been able to discuss here. Some interesting ones that potentially have econometric applications are discussed in Davison, Hinkley, and Young (2003), Hu and Kalbfleisch (2000), Lahiri (2003), Lele (2003), and Shao (2003). Nevertheless, for some models, few or even none of the currently available methods may lead to asymptotically valid inferences. Fewer still may lead to reasonably accurate inferences in finite samples. Consequently, the bootstrap is an active research topic, and the class of models for which the bootstrap can be effectively used is continually growing.

References

Andrews, D. W. K. (2002). "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators," *Econometrica*, 70, 119–162.

Andrews, D. W. K. (2004). "The block-block bootstrap: Improved asymptotic refinements," *Econometrica*, 72, 673–700.

Athreya, K. B. (1987). "Bootstrap of the mean in the infinite variance case," *Annals of Statistics*, 15, 724–731.

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, 83, 687–697.

Berkowitz, J., and L. Kilian (2000). "Recent developments in bootstrapping time series," *Econometric Reviews*, 19, 1–48.

Bühlmann, P. (1997). "Sieve bootstrap for time series," *Bernoulli*, 3, 123–148.

Bühlmann, P. (2002). "Bootstraps for time series," *Statistical Science*, 17, 52–72.

Chang, Y., and J. Y. Park (2003). "A sieve bootstrap for the test of a unit root," *Journal of Time Series Analysis*, 24, 379–400.

Choi, E., and P. Hall (2000). "Bootstrap confidence regions computed from autoregressions of arbitrary order," *Journal of the Royal Statistical Society*, Series B, 62, 461–477.

Davidson, R., and E. Flachaire (2001). "The wild bootstrap, tamed at last," GREQAM Document de Travail 99A32, revised.

Davidson, R., and James G. MacKinnon (1999a). "The size distortion of bootstrap tests," *Econometric Theory*, 15, 361–376.

Davidson, R., and J. G. MacKinnon (1999b). "Bootstrap testing in nonlinear models," *International Economic Review*, 40, 487–508.

Davidson, R., and James G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?," *Econometric Reviews*, 19, 55–68.

Davidson, R., and James G. MacKinnon (2002). "Bootstrap J tests of nonnested linear regression models," *Journal of Econometrics*, 109, 2002, 167–193.

Davidson, R., and James G. MacKinnon (2004). "The power of bootstrap and asymptotic tests," *Journal of Econometrics*, forthcoming.

Davison, A. C., and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*, Cambridge, Cambridge University Press.

Davison, A. C., D. V. Hinkley, and G. A. Young (2003). "Recent developments in bootstrap methodology," *Statistical Science*, 18, 141–157.

DiCiccio, T. J., and B. Efron (1996). "Bootstrap confidence intervals," (with discussion), *Statistical Science*, 11, 189–228.

Dufour, J.-M., L. Khalaf, J.-T. Bernard, and I. Genest (2004). "Simulation-based finite-sample tests for heteroskedasticity and ARCH effects," *Journal of Econometrics*, 122, 317–347.

Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses," *Annals of Mathematical Statistics*, 28, 181–187.

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia, Society for Industrial and Applied Mathematics.

Flachaire, E. (1999). "A better way to bootstrap pairs," *Economics Letters*, 64, 257–262.

Flachaire, E. (2002). "Bootstrapping heteroskedasticity consistent covariance matrix estimator," *Computational Statistics*, 17, 501–506.

Flachaire, E. (2004). "Bootstrapping heteroskedastic regression models: Wild bootstrap vs pairs bootstrap," *Computational Statistics and Data Analysis*, forthcoming.

Freedman, D. A. (1981). "Bootstrapping regression models," *Annals of Statistics*, 9, 1218–1228.

Godfrey, L. G.. (1998). "Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap," *Journal of Econometrics*, 84, 59–74.

Godfrey, L. G., and A. R. Tremayne (2004). "Using the wild bootstrap to implement heteroskedasticity-robust tests for serial correlation in dynamic regression models," *Computational Statistics and Data Analysis*, forthcoming.

Gonçalves, S., and L. Kilian (2004). "Bootstrapping autoregressions with conditional heteroskedasticity of unknown form," *Journal of Econometrics*, 123, 89–120.

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

Hall, P., J. L. Horowitz, and B.-Y. Jing (1995). "On blocking rules for the bootstrap with dependent data," *Biometrika*, 82, 561–574.

Hansen, B. E. (1999). "The grid bootstrap and the autoregressive model," *Review of Economics and Statistics*, 81, 594–607.

Härdle, W., J. L. Horowitz, and J.-P. Kreiss (2003). "Bootstrap methods for time series," *International Statistical Review*, 71, 435–459.

Horowitz, J. L. (2001). "The bootstrap," Ch. 52 in *Handbook of Econometrics*, Vol. 5, ed. J. J. Heckman and E. E. Leamer, Amsterdam, North-Holland.

Horowitz, J. L. (2003). "The bootstrap in econometrics," *Statistical Science*, 18, 211–218.

Hu, F., and J. D. Kalbfleisch (2000). "The estimating function bootstrap," *Canadian Journal of Statistics*, 28, 449–481.

Inoue, A., and L. Kilian (2002). "Bootstrapping smooth functions of slope parameters and innovation variances in VAR(∞) models," *International Economic Review*, 43, 309–331

Jöckel, K.-H. (1986). "Finite sample properties and asymptotic efficiency of Monte Carlo tests," *Annals of Statistics*, 14, 336–347.

Künsch, H. R. (1989). "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217–1241.

Lahiri, P. (2003). "On the impact of the bootstrap in survey sampling and small-area estimation" *Statistical Science*, 18, 199–210.

Lahiri, S. N. (1999). "Theoretical comparisons of block bootstrap methods," *Annals of Statistics*, 27, 386–404.

Lele, S. R. (2003). "Impact of the bootstrap on the estimating functions, *Statistical Science*, 18, 185–190.

Li, H., and G. S. Maddala (1996). "Bootstrapping time series models," (with discussion), *Econometric Reviews*, 15, 115–195.

Liu, R. Y. (1988). "Bootstrap procedures under some non-I.I.D. models, *Annals of Statistics*, 16, 1696–1708.

MacKinnon, J. G. (2002). "Bootstrap inference in econometrics," *Canadian Journal of Economics*, 35, 615–45.

MacKinnon, J. G., and A. A. Smith, Jr. (1998). "Approximate bias correction in econometrics," *Journal of Econometrics*, 85, 205–230.

Mammen, E. (1993). "Bootstrap and wild bootstrap for high dimensional linear models," *Annals of Statistics*, 21, 255–285.

Park, J. Y. (2002). "An invariance principle for sieve bootstrap in time series," *Econometric Theory*, 18, 469–490.

Park, J. Y. (2003). "Bootstrap unit root tests," *Econometrica*, 71, 1845–1895.

Politis, D. N. (2003). "The impact of bootstrap methods on time series analysis," *Statistical Science*, 18, 219–230.

Politis, D. N., and J. P. Romano (1994). "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

Rilstone, P., and M. R. Veall (1996). "Using bootstrapped confidence intervals for improved inferences with seemingly unrelated regression equations," *Econometric Theory*, 12, 569–580

Shao, J. (2003). "Impact of the bootstrap on sample surveys," *Statistical Science*, 18, 191–198.

White, H. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817–838.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis," *Annals of Statistics*, 14, 1261–1295.