

Notes on the Bootstrap

by

Russell Davidson

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

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

`russell@ehess.cnrs-mrs.fr`

These notes were prepared for lectures given at the Université Libre de Bruxelles in April 1998.

Notes on the Bootstrap

1 Preliminaries

We begin by giving definitions of what we mean by models, model parameters, and a number of related things. A **model** is defined as a set of **data-generating processes**, or **DGPs**. A DGP in turn is defined as something that can be simulated on a computer. Although it is possible for a DGP to be deterministic, as might well be the case for simulations of complex nonlinear processes, in econometrics a DGP is usually a **stochastic process**. Often, a DGP can be defined for data sets of arbitrary size, and, indeed, for the purposes of the sort of asymptotic theory usually used in econometrics, it is necessary to be able to consider what happens when the number of observations in a data set tends to infinity.

The essential aspect of a DGP is that it is not defined until enough information has been specified for a computer simulation to be possible. Thus all parameter values must be specified, as must all probability distributions needed to generate the random elements in the simulation. Another aspect of any DGP that is not purely deterministic is that the data it generates will be different for different simulations, because the random elements will be realised differently. Usually, successive simulations will be independent of one another. However, on a computer, a simulation can be reproduced exactly, by resetting the seed of the random number generator to the value it had at the start of the simulation to be reproduced.

When an econometric model is constructed for the purposes of analysing a given data set, the hope is that the model will be large enough to contain a mathematical representation of the real-world economic mechanism that actually did generate the observed data set. If this is so, the model is said to be well or correctly **specified**. If not, then, except in some special cases, no reliable statistical inference can be performed on the basis of estimating the model.

Models and Parametrised Models

Let us denote a model by \mathbb{M} . Then the DGPs in \mathbb{M} are usually characterised, at least partially, by a set of **model parameters**. A **parametrised model** is a model \mathbb{M} coupled with a **parameter-defining mapping**, θ , which takes values in \mathbb{M} and maps them into a **parameter space**, denoted Θ , where $\Theta \subseteq \mathbb{R}^k$ for some positive integer k , which is the number of parameters. Thus each DGP $\mu \in \mathbb{M}$ is associated with a k -dimensional parameter vector $\theta(\mu)$.

The somewhat involved term “parameter-defining mapping” is used because a **parametrisation** would be a map from Θ to \mathbb{M} , whereby one could identify each possible parameter vector with one and only one DGP. A parameter-defining mapping is not required to be one-to-one, and so the usual case is that any given parameter vector corresponds to a whole set of DGPs, and thus constitutes only a partial characterisation.

An easy example of a parameter-defining mapping that is not one-to-one is provided by the simple linear regression model. This model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad E(\mathbf{u} | \mathbf{X}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top | \mathbf{X}) = \sigma^2\mathbf{I}. \quad (1)$$

In this notation, \mathbf{y} is an n -vector each element of which is one of n observations on a dependent variable. Similarly, \mathbf{X} is an $n \times k$ matrix of explanatory variables. Each of its k columns contains the n observations of one of these explanatory variables. The components of the k -vector $\boldsymbol{\beta}$ are model parameters. The n -vector \mathbf{u} contains the random elements in the model, often rather misleadingly referred to as “error terms”. The model specification requires that these error terms have expectation zero and covariance matrix proportional to an identity matrix, conditional on the explanatory variables. The constant of proportionality, which is just the error variance, is another model parameter.

The model (1) is a set of data-generating processes for the dependent variable \mathbf{y} . The other variables appearing in the equation are assumed to be given, and they therefore do not vary from one simulation, or realisation of the DGP, to another, as suggested by the fact that the distribution of the error terms is given conditional on them. Such explanatory variables are said to be (strongly) **exogenous**. Even with the assumption of exogeneity, it is not enough to specify the values of the parameters $\boldsymbol{\beta}$ and σ^2 in order to be able to simulate (1), because the probability distribution from which the error terms are to be drawn has not been completely specified. It is common practice to suppose that the errors are normally distributed, in which case their distribution is fully specified by the first two moments, which are given once σ^2 is known. But other nonnormal distributions are perfectly compatible with such a specification of the first two moments. Thus, for any specification of $\boldsymbol{\beta}$ and σ^2 , there is an infinite number of DGPs that satisfy the requirements of (1).

A **statistic** is any function of the data generated by a DGP. When one says “statistic” in the singular, this normally implies that the statistic is a scalar function of the data. A statistic may, and usually will, depend also on the exogenous variables that figure in the specification of the DGP. It is often convenient, however, to ignore this in notation, and denote a statistic simply as $\tau(\mathbf{y})$. A statistic is by construction a **random variable**, and realisations of it can be obtained by generating a realisation, \mathbf{y}^* say, from the appropriate DGP μ , and computing $\tau^* \equiv \tau(\mathbf{y}^*)$ as the realisation of the statistic.

Clearly, the probability distribution of a given statistic τ depends on that of the data \mathbf{y} , and thus on the DGP μ used to generate \mathbf{y} . A statistic τ is said to be a **pivot**, or to be **pivotal**, for a model \mathbb{M} , if its probability distribution is the same for all DGPs $\mu \in \mathbb{M}$. In other words, the distribution of a pivotal statistic is invariant with respect to the particular DGP of the model that generated it. If we suppose that the DGPs in \mathbb{M} can generate data sets of arbitrary size, it will often be the case that the distribution of a statistic depends on the number of observations, n . The property of pivotalness does not exclude this possibility: All that is needed is that, for *given* sample size n , the distribution of a pivotal statistic does not depend on $\mu \in \mathbb{M}$.

If instead of fixing the sample size, we fix a DGP $\mu \in \mathbb{M}$, we can consider the **asymptotic distribution** of a statistic τ . This is just the limiting distribution of τ as the sample size n tends to infinity; It is, in short, just the asymptotic distribution of the statistic in the usual sense of asymptotic theory. A statistic τ is said to be an **asymptotic pivot** for the model \mathbb{M} if its asymptotic distribution is the same for all $\mu \in \mathbb{M}$. Most test statistics commonly used in econometric practice are asymptotically pivotal, with asymptotic distributions that are standard normal, or chi-squared, or Dickey-Fuller, or the like.

Tests and Monte Carlo Tests

When we use statistics for testing purposes, the model \mathbb{M} corresponds to the hypothesis we wish to test: Every DGP in \mathbb{M} satisfies that hypothesis, called the **null hypothesis**, by construction. In testing situations, we usually also specify an **alternative hypothesis**. It is then required that the distribution of a test statistic τ be different under the null hypothesis and under the alternative, so that the statistic can be used to discriminate between these two hypotheses.

In classical testing, the null hypothesis is a special case of the alternative. The most common example is where a certain parameter is equal to zero under the null but is different from zero under the alternative. Rather than formulating the alternative hypothesis with a point removed from it, it is more convenient to formalise this state of affairs by constructing another model, \mathbb{M}_1 say, to represent the alternative hypothesis, without cutting out the special case. Then the model that corresponds to the null hypothesis, \mathbb{M} , is a subset of \mathbb{M}_1 . The distribution of the statistic τ will usually vary continuously over the whole of \mathbb{M}_1 .

In order to test the null hypothesis, one first computes the value of the statistic from the data \mathbf{y} as $\tau(\mathbf{y})$. For convenience, we will denote this value by $\hat{\tau}$, in order to emphasise that it is a *realisation* of the statistic τ . Next, one can either compare $\hat{\tau}$ with a critical value corresponding to some pre-chosen significance level, and reject the null hypothesis if $\hat{\tau}$ is more extreme than the critical value, or else calculate the marginal significance level, or P value,

corresponding to $\hat{\tau}$. The latter approach is preferable, because knowing the P value associated with a test statistic is more informative than simply knowing whether or not the test statistic exceeds some critical value.

Suppose that $\hat{\tau}$ is computed from data that were generated by a DGP which does indeed belong to \mathbb{M} , and thus satisfies the null hypothesis. We denote this DGP by μ_0 . Under the DGP μ_0 , τ has a well-defined distribution, of course, but, except in very special circumstances, we do not know what this distribution is. We may well know the asymptotic distribution, but that is generally only an approximation. If we knew the exact distribution, then we could compute an ideal P value corresponding to $\hat{\tau}$. Suppose for simplicity that we want to perform a one-tailed test. Then this ideal P value is

$$p(\hat{\tau}) \equiv \Pr_{\mu_0}(\tau \geq \hat{\tau}) = 1 - F_{\mu_0}(\hat{\tau}), \quad (2)$$

where F_{μ_0} is the CDF of τ under μ_0 . This quantity can be called ideal because it really is the probability mass in the tail of the distribution of τ beyond the realized $\hat{\tau}$. By construction, therefore, if our rule is to reject the null hypothesis if $p(\hat{\tau})$ is less than some nominal size α , then the probability of Type I error, that is the probability of falsely rejecting the null when μ_0 is the true DGP, is exactly α . This follows because this probability is

$$\begin{aligned} \Pr_{\mu_0}(p(\hat{\tau}) < \alpha) &= \Pr_{\mu_0}(1 - F_{\mu_0}(\hat{\tau}) < \alpha) \\ &= \Pr_{\mu_0}(F_{\mu_0}(\hat{\tau}) > 1 - \alpha) \\ &= \Pr_{\mu_0}(\hat{\tau} > F_{\mu_0}^{-1}(1 - \alpha)) \\ &= 1 - \Pr_{\mu_0}(\hat{\tau} < F_{\mu_0}^{-1}(1 - \alpha)) \\ &= 1 - F_{\mu_0}(F_{\mu_0}^{-1}(1 - \alpha)) = \alpha, \end{aligned}$$

as required.

In general, the probability in (2) depends both on the sample size n and on the DGP μ_0 . This is not the case if we use asymptotic theory to compute a P value. Denote the CDF of the asymptotic distribution of τ by F_{as} . The asymptotic P value associated with $\hat{\tau}$ is then

$$p_{as}(\hat{\tau}) \equiv 1 - F_{as}(\hat{\tau}). \quad (3)$$

Note that F_{as} , being asymptotic, cannot depend on the sample size n , and, further, cannot depend on the unknown true DGP μ_0 . But this means that, if asymptotic theory is to give the right answer, the ideal $p(\hat{\tau})$ cannot depend either on n or on μ_0 . Only in that case will $p_{as}(\hat{\tau})$ equal $p(\hat{\tau})$.

Since μ_0 is unknown, we cannot compute (2). But if τ is pivotal, we can estimate it by Monte Carlo simulation, because we can simulate using any DGP at all in \mathbb{M} . Since the distribution of τ is constant across \mathbb{M} , it is not

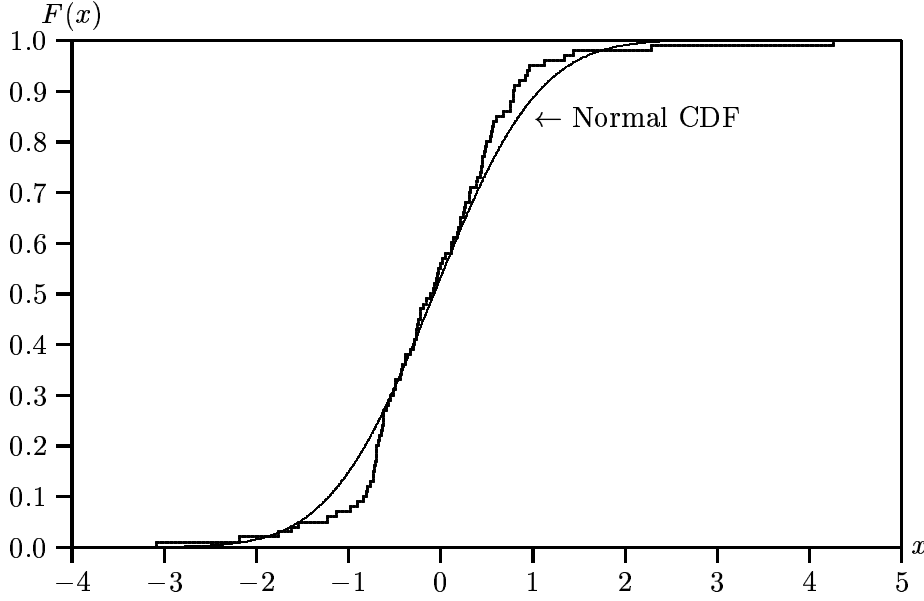


Figure 1. Empirical distribution function based on 100 observations

necessary to know exactly which DGP μ_0 generated the data. In order to estimate a P value by simulation, then, we choose any DGP, μ say, in \mathbb{M} , and make N drawings \mathbf{y}_i^* , $i = 1, \dots, N$, from it. Typically, N will be a rather large number. For each i , we then compute τ_i^* as $\tau(\mathbf{y}_i^*)$. These τ_i^* allow us to define an **empirical distribution function**, or **EDF**, as follows:

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^N I(\tau_i^* \leq x). \quad (4)$$

Here, $I(\cdot)$ denotes an **indicator function**, equal to one if the argument is true, and zero if it is false. Thus $\hat{F}(x)$ is just the proportion of the N realisations for which the simulated value of τ was less than or equal to x . Thus the EDF is a step function, the height of each step being $1/N$, and the width being equal to the difference between two successive values of τ_i^* when they are sorted in increasing order.

It can be shown easily, by use of the law of large numbers, that $\hat{F}(x)$ converges to the true CDF of τ under μ as $N \rightarrow \infty$. As an illustration, consider Figure 1, in which the EDF for a particular set of 100 observations on a random variable distributed as $N(0, 1)$ is shown, along with the true CDF of that distribution, for comparison.

The EDF $\hat{F}(x)$ allows us to obtain a **simulated P value**, which is a simulation estimate of the ideal P value (2). Replacing the unknown CDF F_{μ_0} by \hat{F} gives us the estimate

$$\hat{p}(\hat{\tau}) = 1 - \hat{F}(\hat{\tau}). \quad (5)$$

By making N large enough, this estimate can be made as accurate as we wish. It is not in fact necessary to construct the EDF explicitly in order to compute (5). From the definition (4), we see that

$$\hat{p}(\hat{\tau}) = \frac{1}{N} \sum_{i=1}^N I(\tau_i^* > \hat{\tau}). \quad (6)$$

In words, the simulated P value is just the proportion of the realisations τ_i^* greater than the value $\hat{\tau}$ computed from the real data.

Using a simulated P value constitutes what is called a **Monte Carlo test**. The idea of a Monte Carlo test is generally attributed to Dwass (1957) and Barnard (1963). If we wish to conduct a test at level α , it is highly desirable to choose N so that $\alpha(N+1)$ is an integer. To see why, suppose that we sort the $N+1$ values, $\hat{\tau}$ from the data and the N simulated statistics τ_i^* , $i = 1, \dots, N$, in decreasing order. The rank r of $\hat{\tau}$ in the sorted set can have $N+1$ possible values, $r = 0, 1, \dots, N$, all of them equally likely under the null hypothesis. Here, r is defined in such a way that there are exactly r simulations for which $\tau_i^* > \hat{\tau}$: If $r = 0$, $\hat{\tau}$ is the largest value in the set, if $r = N$, it is the smallest. Thus the simulated P value $\hat{p}(\hat{\tau})$ is just r/N . The Monte Carlo test rejects if $r/N < \alpha$, that is, if $r < \alpha N$. Under the null, the probability that this inequality will be satisfied is the proportion of the $N+1$ possible values of r that satisfy the inequality. If we denote by $[\alpha N]$ the largest integer that is smaller than αN , it is easy to see that there are exactly $[\alpha N] + 1$ such values of r , namely, $0, 1, \dots, [\alpha N]$. Thus the probability of rejection is $([\alpha N] + 1)/(N + 1)$.

Ideally, we would like the rejection probability to be exactly equal to α . This would imply that

$$\alpha = \frac{[\alpha N] + 1}{N + 1}, \text{ that is, } \alpha(N + 1) = [\alpha N] + 1.$$

This can hold only if $\alpha(N+1)$ is an integer. Conversely, if $\alpha(N+1)$ is an integer, then it follows that $[\alpha N] = \alpha(N+1) - 1$, whence

$$\frac{[\alpha N] + 1}{N + 1} = \frac{\alpha(N + 1) - 1 + 1}{N + 1} = \alpha,$$

as desired. Consider a couple of examples. If N were 100 and α were .05, $\hat{\tau}$ could have 101 possible ranks. If we rejected when its rank was less than $\alpha N = 5$, there would be 5 outcomes that would lead to rejection, $r = 0, 1, 2, 3, 4$, and the probability of rejecting would be $5/101 = .0495$. Notice that, in this case, we would obtain an estimated P value precisely equal to .05 in one case out of 101. It is clear that, when that happens, the null hypothesis should *not* be rejected. Similarly, if N were 101, $\hat{\tau}$ could have

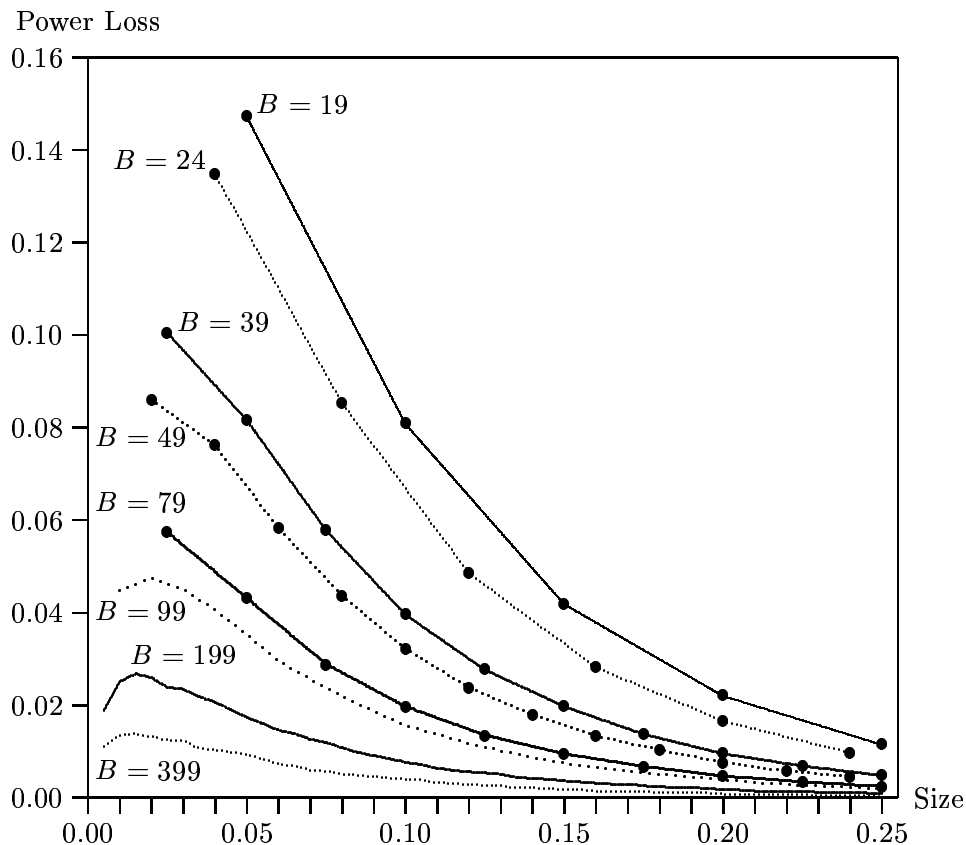


Figure 2. Power Loss from Bootstrapping: $\gamma = 2.0$

102 possible ranks, of which, since $\alpha N = 5.05$, 6 would lead to rejection. Thus the probability of rejecting would be $6/102 = 0.0588$. Observe that the consequences of making $N = 101$ are much more severe than the consequences of making $N = 100$.

We have just seen that $\alpha(N + 1)$ must be an integer if we wish to obtain an exact test. Therefore, for $\alpha = .05$, the smallest possible value of N is 19, and for $\alpha = .01$, the smallest possible value is 99. However, unless computing costs are extraordinarily high, we almost certainly will not want to use these smallest possible values in practice. There are two reasons for this. Firstly, even though using a finite value for N does not affect the *level* of a Monte Carlo test, it does affect the *outcome* of the test. Whether or not we reject will depend on the output of the random number generator we use as well as on the data. Secondly, using a finite value for N reduces the power of the test, often substantially.

To see just how the choice of B affects test power, consider the t statistic for the null hypothesis that $\gamma = 0$ in the very simple model

$$y_t = \gamma + u_t, \quad u_t \sim N(0, 1), \quad t = 1, \dots, 4.$$

Since we know that the exact distribution of the statistic is just $t(3)$ under the null, the ideal P value (2) can be computed. For Monte Carlo tests corresponding to different values of N , the easiest way to proceed is just to make N drawings from the $t(3)$ distribution, and to use equation (6), slightly modified to allow for the fact that this is a two-tailed test, to obtain $\hat{p}(\hat{\tau})$. In Figure 2 there are plotted the results of a simulation experiment, with 400,000 replications, designed to investigate the power loss associated with various values of N . The true value of γ was taken as 2.

The values of N used in the experiment were 19, 24, 39, 49, 79, 99, 199, and 399, because these values yield valid Monte Carlo tests for many commonly encountered values of α . Results are shown only for values of α and N for which the Monte Carlo test has the correct level. In the Figure, the horizontal axis shows test size, and the vertical axis shows the difference between the power of the test based on $N = \infty$ (that is, on the exact P value) and the power of the test based on finite values of N . For $N \leq 79$, the points at which power loss is evaluated are shown as balls.

2 Bootstrap Tests

The name **bootstrap** is based on the same metaphor as that used in computer terminology to refer to the startup process by which a computer configures itself for subsequent use. “Pulling oneself up by one’s own bootstraps” is a rather peculiarly American usage, probably originally meant to signify the impossible feat of propelling one’s body into the air by pulling up on one’s shoelaces, or “bootstraps”, in an earlier idiom. The idea seems to be that a computer gets itself going all by itself when it is turned on, without outside aid. The metaphor, like all those – to Europeans – incomprehensible American baseball metaphors, has absolutely no redeeming literary merit, but, if nothing else, it serves to recycle a word, “bootstrap”, that would otherwise disappear from the language as obsolete. It is probably best to treat the word as new, forgetting that it once meant something else.

The original bootstrap idea was that a sort of Monte Carlo experiment could be performed in which the error terms or other random quantities are drawn, not from a distribution specified by some model, such as the normal distribution, but rather from the empirical distribution function of their sample counterparts. Obtaining artificial samples in this way is a special case of what is called **resampling**; see Efron (1979). Another well-known resampling technique is called the **jackknife**, by use of another ridiculous metaphor, this time for something that has shown itself to be useful. It will be useful at this point to interrupt our development of the bootstrap to discuss what is meant by resampling, and why it might be interesting.

Resampling

Let us begin with a simple case, in which we have a sample of n observations that are supposed to be IID drawings from some unknown distribution. As above, we denote this sample by the n -vector \mathbf{y} . Usually, we will be interested in some statistic $\tau(\mathbf{y})$. As for a Monte Carlo test, we wish to obtain a simulation-based estimate of the distribution of τ , but we do not know what distribution to use in order to generate the simulated data vectors \mathbf{y}_i^* . Instead, we draw the \mathbf{y}_i^* from the **empirical distribution** of the data, that is, the discrete distribution characterised by the EDF of the n elements of the vector \mathbf{y} that constitutes our original data set:

$$\hat{F}(x) = \frac{1}{n} \sum_{t=1}^n I(y_t \leq x).$$

Here, y_t is a typical element of \mathbf{y} . It will be convenient to denote by $\hat{\mu}$ the DGP that generates drawings from this empirical distribution.

In order to generate data using $\hat{\mu}$, we can employ the procedure known as resampling. For a data set of size n , this involves making n successive drawings from the set $\{y_t\}_{t=1}^n$ *with replacement*. It is necessary to replace elements so that the successive drawings are mutually independent drawings from the same distribution. Data generated in this way constitute what is called a **bootstrap sample**. Each bootstrap sample will contain some of the original n observations more than once, and others of them not at all, in a completely random order. Drawing a bootstrap sample is very easy. Let y_j^* denote the j^{th} observation of a bootstrap sample. To obtain y_j^* , we generate a random integer k that takes on the values $1, \dots, n$ with equal probability, and then set y_j^* to y_k . Repeating this operation n times yields a complete bootstrap sample \mathbf{y}^* . We can then calculate the **bootstrap statistic** τ^* as $\tau(\mathbf{y}^*)$.

By making a large number of drawings from $\hat{\mu}$, we can compute a **bootstrap P value** by the formula (6). In the case of a Monte Carlo test, the simulated P value tends to the ideal P value (2) when $N \rightarrow \infty$. However, this is not true for a bootstrap P value, because we use $\hat{\mu}$ rather than the true DGP μ_0 , and, even if τ is a pivot for a given model, we cannot expect that its distribution will be the same under a discrete DGP like $\hat{\mu}$ as it would be under μ_0 . Incidentally, it is conventional to use B rather than N to denote the number of bootstrap samples used in an experiment, and we will henceforth use this conventional notation.

Although a bootstrap P value differs in general from the ideal P value (2) even when $B \rightarrow \infty$, it can be justified asymptotically, in the limit when the sample size, n , tends to infinity. In that case, if the original data \mathbf{y} are generated by the DGP μ_0 , the EDF of the elements of \mathbf{y} tends to the true CDF, F say, of those elements, by the law of large numbers. Consider now a bootstrap sample \mathbf{y}^* of m elements, for arbitrary m . Each element of \mathbf{y}^* is drawn

from a distribution that approaches F as $n \rightarrow \infty$. Thus the distribution of a bootstrapped statistic τ^* approaches the distribution it would have if the data were generated by μ_0 .

Normally we set $m = n$, because we calculate the statistic τ using the n -vector \mathbf{y} . Thus m tends to infinity with n , and the bootstrap distribution of the τ tends to the distribution with CDF F_{as} . Since the distribution of τ itself also tends to this distribution, we conclude that, in the limit as $n \rightarrow \infty$, the distributions of the statistic under μ_0 and $\hat{\mu}$ coincide, so that the bootstrap P value coincides with the ideal one. This asymptotic argument applies to any statistic τ under weak smoothness conditions: We obviously require that the distribution of τ varies smoothly with the distribution of the elements of \mathbf{y} .

Now suppose that we wish to work in the context of a model \mathbb{M} . Provided a statistic τ satisfies an appropriate smoothness condition, we have just seen that the bootstrap P value is asymptotically equal to the ideal one, whether or not τ is a pivot for \mathbb{M} . This remark is the fundamental justification for the use of the bootstrap, for all sorts of purposes, not just for hypothesis testing. The original motivation seems to have been a desire for **robustness** of inference: Since μ_0 is unknown, and since any model we study may well not contain μ_0 , it may well be safer just to estimate μ_0 by a DGP $\hat{\mu}$ based on resampling if we wish to obtain robust estimates of some aspect, a standard error for instance, of the distribution of a statistic or estimator. It appears that, under a weak regularity condition, this is just fine asymptotically. Note that any quantity that has to be estimated within the context of a model \mathbb{M} cannot be pivotal, for otherwise its value, being the same for all DGPs in the model, would be known.

The IID context is of course very restrictive. To give at least an indication of how resampling can be used in more general contexts, let us consider a possibly nonlinear regression model

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad t = 1, \dots, n, \quad (7)$$

where any variables on which $x_t(\boldsymbol{\beta})$ depends are assumed to be fixed or at least independent of all the u_t . If these are assumed to be IID, the natural approach is to bootstrap by resampling the residuals. With this approach, one first estimates the model (7) by NLS, so as to obtain parameter estimates $\hat{\boldsymbol{\beta}}$ and residuals \hat{u}_t , $t = 1, \dots, n$, and then sets up the bootstrap DGP as

$$y_t^* = x_t(\hat{\boldsymbol{\beta}}) + u_t^*, \quad t = 1, \dots, n, \quad (8)$$

where the u_t^* are drawn randomly with replacement from the set $\hat{u}_1, \dots, \hat{u}_n$. If $x_t(\boldsymbol{\beta})$ depends on the lagged dependent variable, y_{t-1} , this approach can still be employed, but then the lags y_{t-1}^* must be used in (8) in place of the y_{t-1}

from the original data. Thus each y_t^* will be computed recursively, since it depends on y_{t-1}^* . This in turn poses the problem of how to start the recursion. Since one often omits the first observation in running regressions with a lagged dependent variable, one solution is just to use the first observation, y_1 , from the original data in each bootstrap sample. Another might be to compute the stationary distribution of the y_t and to draw from that. However, it may be the case that no stationary distribution exists, or, if it does, it may depend on the model parameters. In the latter case, $\hat{\beta}$ would be used.

Parametric Bootstrap Methods

As more has been learnt about the bootstrap over the years, the focus has tended to move away from robustness towards accuracy. Monte Carlo tests can provide exact inference, up to simulation error, when exact pivots are used, and it is plausible to think that the bootstrap, which is also a simulation method, should do the same. This is in fact the case if the **parametric bootstrap** is used.

The parametric bootstrap makes no use of resampling in setting up the bootstrap DGP. Rather, a fully parametric model (\mathbb{M}, θ) is used, with a one-to-one parameter-defining mapping θ . The first step is to obtain estimates $\hat{\theta}$ of the model parameters, by least squares, or maximum likelihood, or some other appropriate estimation technique, and to compute the value $\hat{\tau}$ of the statistic, which may or may not be pivotal, on which one wishes to base inference. Next the bootstrap DGP $\hat{\mu}$ can be set up as the DGP characterised by the estimated parameters $\hat{\theta}$. This DGP is uniquely determined because we assumed that θ is one to one. A large number B of bootstrap samples are generated from $\hat{\mu}$; for each of these a bootstrap statistic τ^* is computed; a bootstrap P value is computed using (6).

If τ is exactly pivotal, then the distribution of the τ^* will be the same as that of τ itself under any DGP in \mathbb{M} , because the bootstrap DGP $\hat{\mu}$ belongs to \mathbb{M} by construction. There is thus no difference, other than a terminological one, between a Monte Carlo test and a bootstrap test in this case. Although the case of an exact pivot is rather special, it has numerous applications in econometrics. For example, in univariate linear regression models with normal errors and regressors that can be treated as fixed, any specification test that depends only on the residuals and the regressors will be pivotal. This includes a great many commonly used diagnostic tests. Thus, provided the assumption of normal error terms is maintained, all of these commonly used tests can be made exact by using the parametric bootstrap.

Suppose next that τ is only asymptotically pivotal, without being an exact pivot. The parametric bootstrap can still be applied, using the method outlined above. We will see that doing so leads to inference which, unlike that provided by exact pivots, is not exact, but is nevertheless better, in a sense to be defined, than inference based on an asymptotic P value. Intuitively,

this is because the bootstrap DGP generates bootstrap statistics for the same sample size as that of the original data, and so is in error only because the parameter estimates $\hat{\boldsymbol{\theta}}$ are not equal to the true parameters. This remaining error is small for two separate reasons: First, the estimates, being consistent, are close to the true parameters, and, second, since the distribution of an asymptotic pivot does not vary as a function of the model parameters in the limit of large sample size, it varies only slightly in finite samples. Thus the error in the bootstrap distribution results from the effect, small even for large parameter differences, of the small estimation error. This argument will be presented more formally later, and specific examples will be given to show that the parametric bootstrap usually behaves very well indeed in standard econometric applications.

An important example of the parametric bootstrap is provided by the nonlinear regression model, which is a fully specified model if we specify the error distribution. We take for our null hypothesis the model

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (9)$$

and our alternative hypothesis will be represented by the model

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{u}, \quad (10)$$

where we assume that the $\mathbf{x}(\boldsymbol{\beta})$ of (9) equals $\mathbf{x}(\boldsymbol{\beta}, \mathbf{0})$ with the \mathbf{x} of (10), so that the null is nested in the alternative. We assume that $\boldsymbol{\beta}$ is a k_1 -vector, that $\boldsymbol{\gamma}$ is a k_2 -vector, and that $k = k_1 + k_2$.

There are many test statistics that could be used to test (9) against (10), and all of them can be bootstrapped. One that is simple to describe is the pseudo- F test: one estimates both (9) and (10) by NLS, and saves the sums of squared residuals from these as SSR_0 and SSR_1 respectively. The test statistic is then

$$\hat{F} = \frac{SSR_0 - SSR_1}{SSR_1} \frac{n - k}{k_2}, \quad (11)$$

where as usual n is the sample size. The realized value of statistic \hat{F} can then be compared against the $F(k_2, n - k)$ distribution, or else it can be multiplied by k_2 and then compared against the asymptotic $\chi^2(k_2)$ distribution. Either of these procedures leads to a P value with only an asymptotic justification, and, in general, neither yields exact inference.

A possible choice for the bootstrap DGP is:

$$\mathbf{y} = \mathbf{x}(\tilde{\boldsymbol{\beta}}, \mathbf{0}) + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \tilde{\sigma}^2 \mathbf{I}), \quad (12)$$

where $\tilde{\boldsymbol{\beta}}$ is the vector of estimates obtained by estimating (9) with the restrictions $\boldsymbol{\gamma} = \mathbf{0}$ imposed, and $\tilde{\sigma}^2$ is restricted error variance estimate. There

are two points to be made about this choice of bootstrap DGP. The first is just that it satisfies the null hypothesis. This is obviously necessary, since any P value at all is based on the distribution of the test statistic under the null. The second is that it uses the *restricted* estimates $\tilde{\beta}$. We would have a bootstrap DGP satisfying the null if we were to use the unrestricted estimates $\hat{\beta}$ obtained by estimating (10), simply setting to zero the parameters γ that appear only under the alternative. This second bootstrap DGP is in fact valid, but it can be shown that it leads to somewhat less good inference than the choice (12).

The next step in the bootstrapping is to draw B bootstrap samples from the bootstrap DGP $\hat{\mu}$ given by (12). For this, we may first evaluate the vector of regression functions $\mathbf{x}(\tilde{\beta}, \mathbf{0})$, using the values of the exogenous variables observed in the original data, and the estimates $\tilde{\beta}$. Next, for $i = 1, \dots, B$, we draw the vector \mathbf{u}_i^* of error terms for the i^{th} bootstrap sample, by using a random number generator to provide n independent $N(0, 1)$ random variables, and then multiplying each of these by $\tilde{\sigma}$ so as to give them variance $\tilde{\sigma}^2$. The “data” \mathbf{y}_i^* for the i^{th} bootstrap sample are then computed by the formula:

$$\mathbf{y}_i^* = \mathbf{x}(\tilde{\beta}, \mathbf{0}) + \mathbf{u}_i^*.$$

Next, for each $i = 1, \dots, B$ we repeat whatever testing procedure we used with the original data. For the pseudo- F test, we run the two nonlinear regressions (9) and (10) using \mathbf{y}_i^* in place of the observed \mathbf{y} . We compute the two sums of squared residuals, and then the bootstrap test statistic F_i^* by formula (11). Finally, the bootstrap P value is computed using (6) with \hat{F} and F_i^* in place of $\hat{\tau}$ and τ_i^* .

The method discussed above should be contrasted with the way in which the bootstrap was used 15 years ago, say. Then the commonest procedure was to use a confidence region. For ease of exposition, we will suppose that γ is just a scalar parameter γ , so that we can argue in terms of confidence intervals rather than regions. The first step would be to estimate, not (9), but (10), so as to obtain the unrestricted estimate $\hat{\gamma}$. Then the bootstrap DGP would almost certainly be based on resampling the residuals from (10), using $\hat{\beta}$ in the regression function. In fact, in some procedures the parameters of the bootstrap DGP are $(\hat{\beta}, \hat{\gamma})$, but this complicates matters unnecessarily, and we will not discuss this further.

The next step would be to obtain a bootstrap estimate of the standard error of $\hat{\gamma}$. For each bootstrap sample, $\hat{\gamma}^*$ would be calculated, and the standard error of the set of bootstrapped statistics, σ^* say, would be used to construct the confidence interval $[\hat{\gamma} - c_\alpha \sigma^*, \hat{\gamma} + c_\alpha \sigma^*]$, where c_α would usually be the critical value of the $N(0, 1)$ distribution appropriate for a significance level of α .

A somewhat more sophisticated approach is the **percentile method**. Here the limits of the bootstrap confidence interval are given in terms of the quantiles

of the distribution of the set of bootstrapped $\hat{\gamma}^*$. The method of using these quantiles to obtain a confidence interval turns out to be surprisingly contorted, and we discuss it no further. Note that both the method based on a bootstrap standard error and the percentile method bootstrap $\hat{\gamma}$, which is by no means a pivotal quantity.

Somewhat more recently, the so-called **percentile- t method** has gained favour. In this method, the bootstrap DGP would use $(\hat{\beta}, \hat{\gamma})$ as parameters. Once more, the implementation of this method is distinctly tricky, and not all authors have carried it off correctly. For a full discussion of the merits and otherwise of these methods, see the introduction to Hall's (1992) classic book, in which much of the more modern theory of the bootstrap is set forth. It is probably fair to say that all these earlier bootstrap methods were harder to implement than the one we used here, and that they give less reliable inference.

3 Bootstrap Refinements

Consider a test based on a statistic τ that is asymptotically pivotal for a model \mathbb{M} . The **asymptotic P value** is given by (3), and it is based on the asymptotic distribution of τ under any DGP $\mu \in \mathbb{M}$. For what follows, we will denote the asymptotic CDF simply as F . At nominal level α , then, the asymptotic test rejects if

$$1 - F(\hat{\tau}) \leq \alpha.$$

We introduce the **P value function**, or **PVF**, as a measure of the true rejection probability:

$$S(\alpha, \mu) \equiv \Pr_{\mu}(1 - F(\tau) \leq \alpha). \quad (13)$$

The notation S signifies that the function gives the (true) *size* of the test at nominal level α under μ . In addition, $S(\cdot, \mu)$ is the CDF of the asymptotic P value under μ . The difference between $S(\alpha, \mu)$ and α will be referred to as the **P value discrepancy function**. It is implicitly defined by the equation

$$S(\alpha, \mu) = \alpha + n^{-l/2} s(\alpha, \mu), \quad (14)$$

where $l \geq 1$ is defined so that $s(\alpha, \mu)$ will be $O(1)$. Results discussed in Hall (1992) imply that the value of l will be different in different cases. In the case of a one-sided test based on an asymptotically $N(0, 1)$ statistic, $l = 1$. In the case of a two-sided test based on an asymptotically $N(0, 1)$ statistic and the case of an asymptotically χ^2 statistic, $l = 2$. Where no ambiguity is possible, we will often refer to the function s itself as the P value discrepancy function.

If each DGP $\mu \in \mathbb{M}$ is fully characterized by a parameter vector θ , the P value function can be written as $S(\alpha, \theta)$. Suppose that a data set actually generated by θ_0 yields estimates $\hat{\theta}$. Then the parametric bootstrap DGP is characterized

by $\hat{\boldsymbol{\theta}}$. The probability of rejection by the asymptotic test can be calculated under the DGPs corresponding to both $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. From (13) and (14), the difference between the two probabilities is

$$S(\alpha, \hat{\boldsymbol{\theta}}) - S(\alpha, \boldsymbol{\theta}_0) = n^{-l/2} (s(\alpha, \hat{\boldsymbol{\theta}}) - s(\alpha, \boldsymbol{\theta}_0)).$$

If s is differentiable, it can be Taylor expanded around $\boldsymbol{\theta}_0$ to obtain

$$S(\alpha, \hat{\boldsymbol{\theta}}) - S(\alpha, \boldsymbol{\theta}_0) \stackrel{a}{=} n^{-l/2} \mathbf{s}_{\boldsymbol{\theta}}^{\top}(\alpha, \boldsymbol{\theta}_0) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad (15)$$

where $\mathbf{s}_{\boldsymbol{\theta}}(\alpha, \boldsymbol{\theta}_0)$ is the vector of first derivatives of $s(\alpha, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}_0$. If $\hat{\boldsymbol{\theta}}$ is root- n consistent, the quantity on the right-hand side of (15) is of order $n^{-(l+1)/2}$. Thus the bootstrap approximation to the distribution of τ is in error only at order $n^{-(l+1)/2}$, better than the error of the asymptotic uniform distribution by a factor of $n^{-1/2}$. To the best of my knowledge, this analysis first appeared in Beran (1988).

The argument of the previous paragraph establishes that, when used with an asymptotic pivot, the parametric bootstrap provides an **asymptotic refinement** relative to the asymptotic test. Note that, if the statistic were not asymptotically pivotal, the leading term in (14) would not be independent of μ , and so the leading term in (15) would not have a factor of $n^{-l/2}$. The bootstrap P value would thus be in error at order $n^{-1/2}$, no better than the asymptotic test.

The information contained in the function $S(\alpha, \mu)$ is also provided by the **critical value function**, or **CVF**, $Q(\alpha, \mu)$, defined implicitly by the equation

$$\Pr_{\mu}(\tau \geq Q(\alpha, \mu)) = \alpha. \quad (16)$$

$Q(\alpha, \mu)$ is thus the true level- α critical value for τ if the DGP is μ . It is easy to see that the relation between the PVF $S(\alpha, \mu)$ and the CVF $Q(\alpha, \mu)$ is

$$S(1 - F(Q(\alpha, \mu)), \mu) = \alpha, \quad (17)$$

and that the rejection region for the bootstrap test of nominal level α is

$$\tau \geq Q(\alpha, \hat{\mu}), \quad (18)$$

so that the rejection probability of the bootstrap test under any DGP μ is the probability of the event (18) under μ .

In the parametric case, the DGP $\mu \in \mathbb{M}$ is completely characterised by the parameter vector $\boldsymbol{\theta}$, and so we can, at least in principle, graph the CVF as a function of $\boldsymbol{\theta}$. As an illustration, Figure 3 shows the CVF, for $\alpha = .05$, for a hypothetical nonpivotal test statistic, asymptotically distributed as the absolute value of $N(0, 1)$, for a model the DGPs of which are characterized

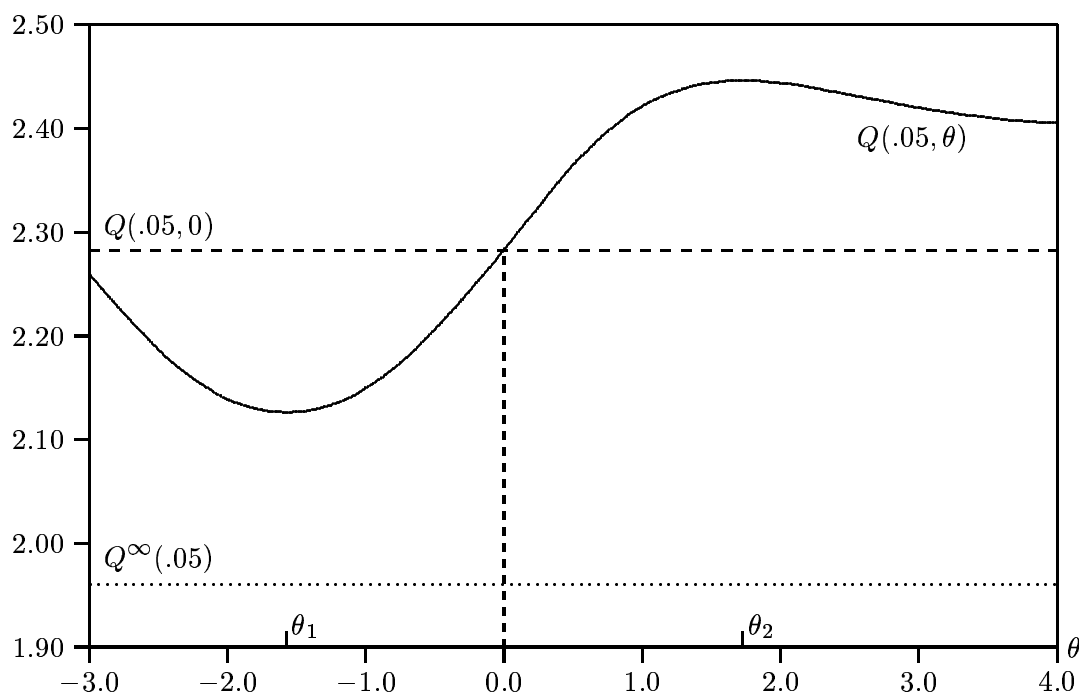


Figure 3. A Critical Value Function

by a single parameter θ . It follows from (18) that the rejection region of the bootstrap test can be defined in the space of $\hat{\theta}$ and τ , as the region above the graph of the CVF.

The rectangle above the horizontal line marked $Q(.05, 0)$ in the figure shows all $(\hat{\tau}, \hat{\theta})$ pairs that would lead to rejection at the .05 level when $\theta_0 = 0$ if the ideal P value (2) could be used. In contrast, the area above the CVF shows all pairs that actually *will* lead to rejection using a bootstrap test. The effect of the difference between the two rejection regions on the performance of the bootstrap test depends on the joint distribution of τ and $\hat{\theta}$. For comparison, the rectangle above the dotted line shows all pairs that lead to rejection with the asymptotic critical value $Q^\infty(.05) = 1.96$. Clearly, the bootstrap test will work much better than the asymptotic test.

From Figure 3, we see that, when $\theta_0 = 0$, the bootstrap test may either overreject or underreject. For values of $\hat{\theta}$ reasonably near 0, it will overreject when $\hat{\theta} < 0$. For those values of $\hat{\theta}$, the CVF is below $Q(.05, 0)$, and the bootstrap critical value will consequently be too small. Similarly, the bootstrap test will underreject whenever $\hat{\theta} > 0$. If $\hat{\theta}$ is approximately unbiased and not very variable, these two types of errors should tend to offset each other, since the CVF is approximately linear near $\theta = 0$. In contrast, near the minimum θ_1 and the maximum θ_2 , all the errors will be of the same sign. Thus we would

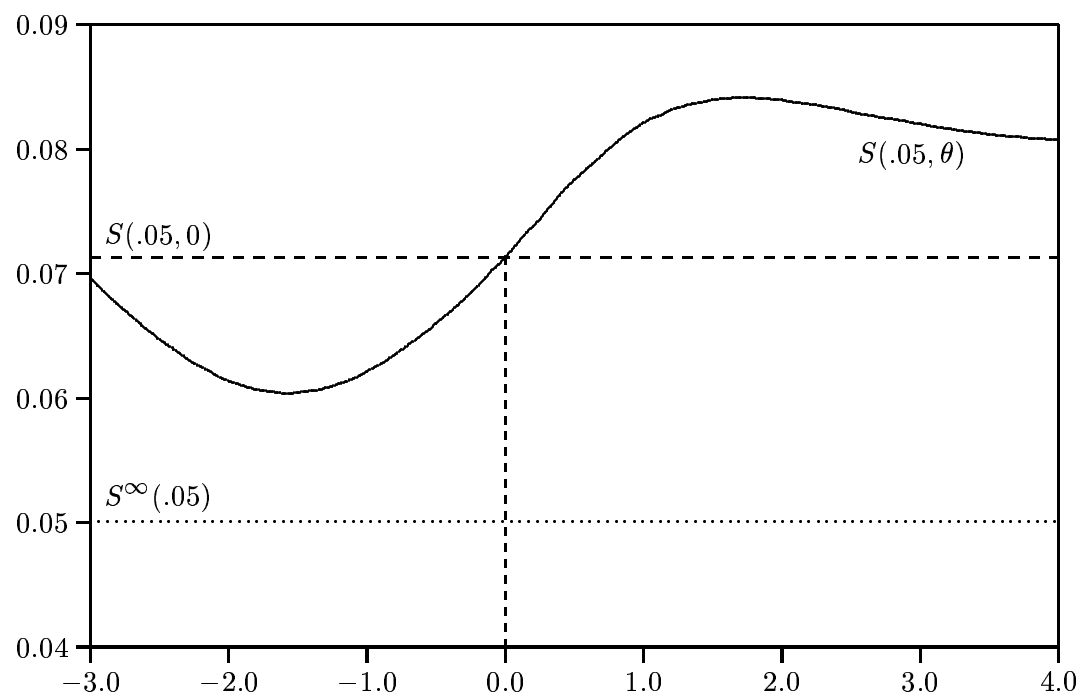


Figure 4. A P Value Function

expect the bootstrap test to underreject when $\theta_0 = \theta_1$ and to overreject when $\theta_0 = \theta_2$.

Figure 4 graphs the PVF $S(.05, \theta)$ for exactly the same one-parameter case as the CVF in Figure 3. Both functions evidently convey essentially the same information.

It is convenient for our analysis to work, not in terms of a statistic which rejects when its value is too large, but rather in terms of the asymptotic P value $1 - F(\tau)$, which we henceforth denote simply by τ . The sign of the inequality in (16) must be changed, because one rejects when a P value is less, rather than greater, than a given value. For statistics τ that are asymptotic P values, the CVF is defined by

$$\Pr_{\mu}(\tau \leq Q(\alpha, \mu)) = \alpha$$

rather than by (16). Clearly, $Q(\alpha, \mu)$ is now the α quantile of τ under μ . Similarly, (13) simplifies to $S(\alpha, \mu) = \Pr_{\mu}(\tau \leq \alpha)$, and (17) reduces to

$$S(Q(\alpha, \mu), \mu) = \alpha, \quad (19)$$

from which it is clear that $Q(\alpha, \mu)$ is the inverse function of $S(\alpha, \mu)$ for given μ . Since both S and Q are increasing in their first arguments, (19) implies that

$Q(S(\alpha, \mu), \mu) = \alpha$. Analogously to (14), we have

$$Q(\alpha, \mu) = \alpha + n^{-l/2}q(\alpha, \mu), \quad (20)$$

with the function q of order unity. The l in (20) is the same as the l in (14).

Equations like (14) and (20), expressed in terms of negative powers of n , do not look very much like the Edgeworth expansions about the standard normal density used in, among others, Hall (1992). They are, however, fundamentally similar, the differences being due to our use of statistics that are asymptotically uniform on $[0, 1]$, rather than standard normal. The uniform density turns out to be better adapted to the study of size distortion.

The bootstrap critical value for τ at nominal level α is $Q(\alpha, \hat{\mu})$, a random variable that is asymptotically nonrandom and equal to α . For given α , it is convenient to define a new random variable γ , of order unity as $n \rightarrow \infty$, as follows:

$$Q(\alpha, \hat{\mu}) = Q(\alpha, \mu_0) + n^{-k/2}\gamma, \quad (21)$$

where k is chosen to make (21) true. For the parametric bootstrap with root- n consistent estimates, Beran's argument in the last section implies that $k = l + 1$. In fact, this is also true for the nonparametric bootstrap whenever Hall's Edgeworth expansion theory applies. In both these cases, (15) shows that $S(\alpha, \hat{\mu}) - S(\alpha, \mu_0)$ is $O(n^{-(l+1)/2})$, as must also be $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$. Thus, since $l \geq 1$, we can be confident that $k \geq 2$ in most cases of interest. Clearly, what is needed is that we should be able to write (20) not only for $\mu \in \mathbb{M}$, but also for the discrete-valued DGPs $\hat{\mu}$ that are used as nonparametric bootstrap distributions. Then, provided that $\hat{\mu} - \mu_0 = O(n^{-1/2})$, $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$ will be $O(n^{-(l+1)/2})$ under standard regularity conditions.

In order to calculate the rejection probability of the bootstrap test under the DGP μ_0 , we need the joint distribution under μ_0 of τ and γ . The PVF $S(\cdot, \mu_0)$ is the marginal CDF of τ . For the joint distribution, we also need the distribution of γ conditional on τ . Thus let $g(\gamma | \tau)$ be the conditional density under μ_0 . With this specification, we can compute the bootstrap rejection probability as the probability under μ_0 that $\tau \leq Q(\alpha, \hat{\mu})$. By (21), this is

$$\int_{-\infty}^{\infty} d\gamma \int_0^{Q+n^{-k/2}\gamma} dS(\tau) g(\gamma | \tau), \quad (22)$$

where, for ease of notation, we have set $Q = Q(\alpha, \mu_0)$ and $S(\tau) = S(\tau, \mu_0)$.

The integral over τ in (22) can be split into two parts, as follows:

$$\int_0^Q dS(\tau) \int_{-\infty}^{\infty} d\gamma g(\gamma | \tau) + \int_{-\infty}^{\infty} d\gamma \int_0^{n^{-k/2}\gamma} d\tau S'(Q + \tau) g(\gamma | Q + \tau), \quad (23)$$

where S' is the derivative of $S(\tau)$. Because g is a density, the integral over γ in the first term of (23) equals 1, and so the whole first term equals α , by (19). The size distortion, or error in rejection probability, is therefore given by the second term in (23). By (14), $S'(Q + \tau) = 1 + O(n^{-l/2})$. Thus, if g is smooth enough, we may write the second term of (23) as

$$\begin{aligned} & n^{-k/2} \int_{-\infty}^{\infty} d\gamma \gamma (g(\gamma | \alpha) + O(n^{-k/2})) (1 + O(n^{-l/2})) \\ &= n^{-k/2} \int_{-\infty}^{\infty} d\gamma \gamma g(\gamma | \alpha) + O(n^{-(k+l)/2}). \end{aligned} \quad (24)$$

Expression (24) is the size distortion of the bootstrap test. The first term is $O(n^{-k/2})$, in accord with Beran's (1988) analysis for asymptotic pivots. We now go beyond his analysis by exploiting the explicit expression in (24) for the leading-order distortion.

The leading-order term in (24) is the expectation, conditional on $\tau = \alpha$, of $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$. Thus it is the bias, conditional on $\tau = \alpha$, of the bootstrap critical value for nominal level α . When this bias is nonzero and of order $n^{-k/2}$, it is responsible for the leading-order size distortion of the bootstrap test. But if it is zero or of lower order, the size distortion is of lower order, because, in that case, those $\hat{\mu}$ which overestimate the critical value will on average be balanced to leading order by those $\hat{\mu}$ which underestimate it. Specifically, if

$$\int_{-\infty}^{\infty} d\gamma \gamma g(\gamma | \alpha) = O(n^{-i/2}),$$

for $i \leq k$, the distortion (24) is of order $n^{-(k+i)/2}$.

The actual value of k in specific testing situations can often be found in the existing literature on bootstrap confidence intervals, because the limits of these intervals are defined in terms of quantiles of the bootstrap distribution, which is precisely what $Q(\alpha, \hat{\mu})$ is. For instance, in Section 3.6 of Hall (1992), it is shown that, for a symmetric confidence interval based on an asymptotically standard normal statistic, the exact and bootstrap critical values differ only at order $n^{-3/2}$. In our terms, $k = 3$. Hall goes on to show that the coverage error of a bootstrap confidence interval is of still lower order, namely, n^{-2} in general. Examination of Hall's derivation of this result reveals that the additional refinement is due, as (24) would suggest, precisely to the fact that the *bias*, or expectation of the difference between the true and bootstrap critical values, is of lower order than the difference itself.

In fact, the interpretation of the leading-order term of (24) as a conditional expectation makes it obvious why it vanishes for a symmetric two-tailed test. The condition that $\tau = \alpha$ corresponds to two possibilities for the underlying asymptotically standard normal statistic: it may be equal to either the

positive critical value or its negative. Under the null, because the test is symmetric, these events are equally probable to leading order. If γ and the asymptotically standard normal statistic have an approximate bivariate normal distribution, as they do in Hall's demonstration, then the expectations of γ conditional on the two critical values, positive and negative, are equal and opposite in sign. Thus the expectation of γ conditional on $\tau = \alpha$ vanishes to leading order.

A Further Refinement

In general, without any requirement of a symmetric two-sided test, the first term in (24) will vanish to leading order if a further condition is satisfied. This condition is that γ and τ should be asymptotically independent.

The asymptotic independence of γ and τ can often be achieved by using the fact that parameter estimates under the null are asymptotically independent of the statistics associated with tests of that null in a wide variety of circumstances. If $\hat{\theta}$ is an extremum estimator that satisfies first-order conditions in the interior of the parameter space of the null hypothesis, the vector $n^{1/2}(\hat{\theta} - \theta_0)$ will be asymptotically independent of any classical test statistic. For the case of the classical test statistics based on maximum likelihood estimation, a detailed proof of this may be found in Davidson and MacKinnon (1987). The proof can be extended in regular cases to NLS, GMM, and other forms of extremum estimation.

Thus, for the parametric bootstrap, the condition of asymptotic independence of τ and γ will always be satisfied, provided the parameters are estimated under the null and have the usual asymptotic properties. This is because $Q(\alpha, \hat{\mu})$, and hence γ , is simply a function of the vector of parameter estimates, and is thus asymptotically independent of the test statistic τ .

Asymptotic independence can often be achieved with little trouble for the non-parametric bootstrap as well. As one example, consider bootstrapping a test statistic in a linear regression model that includes a constant term by resampling from the residuals. Here, in order to achieve asymptotic independence, the bootstrap DGP would be based on estimating the model under the null. The bootstrap regression function would be given by the fitted values, which, as before, are asymptotically independent of any classical test statistic, and the bootstrap error terms would be independent drawings from the empirical distribution of the residuals, u_t . Now consider a t statistic on a variable not included under the null. Asymptotically, it will be a linear combination of the residuals, and so not independent of the bootstrap distribution. To leading order asymptotically, it will have the form

$$\tau \equiv n^{-1/2} \sum_{t=1}^n x_t u_t, \quad \text{where} \quad \sum_{t=1}^n x_t = 0 \quad \text{and} \quad \sum_{t=1}^n x_t^2 = 1.$$

The quantiles of the bootstrap distribution are determined by the empirical distribution of the residuals. Consider the asymptotic covariance of τ and the empirical distribution function evaluated at z . It is

$$\lim_{n \rightarrow \infty} E \left(n^{-1/2} \sum_{t=1}^n x_t u_t, n^{-1/2} \sum_{t=1}^n (I(u_t < z) - E(I(u_t < z))) \right), \quad (25)$$

Since the residuals are asymptotically IID, (25) becomes

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n x_t E \left(u_t (I(u_t < z) - E(I(u_t < z))) \right) = m \lim_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n x_t = 0,$$

where $m = E(u_t (I(u_t < z) - E(I(u_t < z))))$ is independent of t . Since both τ and the empirical CDF are asymptotically normal, they are asymptotically independent because their asymptotic covariance is zero. This kind of result can be obtained much more generally, of course.

Suppose now that γ and τ are asymptotically independent. Then the conditional density of γ becomes

$$g(\gamma | \tau) = g(\gamma) (1 + n^{-j/2} f(\gamma, \tau)),$$

where $g(\gamma)$ is the asymptotic marginal density of γ , and $j \geq 1$ is chosen so that $f(\gamma, \tau)$ is of order unity as $n \rightarrow \infty$.

For any valid bootstrap procedure, that is, one that satisfies the weak smoothness condition mentioned earlier, $q(\alpha, \hat{\mu})$ must be a consistent estimator of $q(\alpha, \mu_0)$, so that

$$\int_{-\infty}^{\infty} d\gamma \gamma g(\gamma) = 0,$$

for otherwise $q(\alpha, \hat{\mu})$ would be asymptotically biased. Then (24) becomes

$$n^{-(k+j)/2} \int_{-\infty}^{\infty} d\gamma \gamma g(\gamma) f(\gamma, \alpha) + O(n^{-(k+j+1)/2}). \quad (26)$$

The interpretation of (26) is the same as that of (24). The first term is the bias, conditional on $\tau = \alpha$, of the bootstrap critical value at nominal level α . When $k = 2$ and $j = 1$, this term will be $O(n^{-3/2})$.

The various refinements discussed should by now make it clear why we chose to implement the parametric bootstrap of a regression model in the way we did. First, by bootstrapping the pseudo- F test, which is an asymptotic pivot under the null hypothesis, we ensure that the first refinement will be available. Second, by setting up the bootstrap DGP so as to depend only on

the parameter estimates under the null, we guarantee that the second refinement will also be available, because the pseudo- F statistic is asymptotically independent of those parameter estimates.

Before concluding this section, we may note that the analysis above is to be found in Davidson and MacKinnon (1996a). A similar analysis can be applied to the power of bootstrap tests, and this can be found in Davidson and MacKinnon (1996b). The principal result of the power analysis is that, on a proper size-corrected basis, the power of a bootstrap test differs from that of the corresponding asymptotic test at the same order as that of the bootstrap test's size distortion.

4 Some Examples

Although the bootstrap almost always leads to very substantial improvements in the reliability of commonly used econometric testing procedures, it is entirely possible to misuse the bootstrap and end up with results no better than those provided by asymptotic theory. In this section, we will look at a few examples that illustrate some potential traps, and how to avoid them.

Bootstrap Tests for Serial Correlation

The problem of testing for serial correlation in regression models has been of central concern to econometricians for roughly half a century. For simplicity, we will restrict our attention to univariate, linear models of the form

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{Y}_t\boldsymbol{\delta} + u_t, \quad u_t = \sum_{l=1}^r \rho_l u_{t-l} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma^2), \quad (27)$$

where \mathbf{X}_t is a $k \times 1$ vector of exogenous regressors that may be treated as fixed, \mathbf{Y}_t is an $m \times 1$ vector of lagged values of the dependent variable y_t , and $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are, respectively, a k -vector and an m -vector of parameters. The normality assumption is essential for some of our results, but not for most of them.

One widely used way to test the null hypothesis that all the ρ_l are zero is based on the Gauss-Newton regression. First, estimate the model (27) under the null hypothesis so as to obtain estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$, and residuals $\hat{u}_t \equiv y_t - \mathbf{X}_t\hat{\boldsymbol{\beta}} - \mathbf{Y}_t\hat{\boldsymbol{\delta}}$, and then run the regression

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{Y}_t\boldsymbol{\delta} + \sum_{l=1}^r \rho_l \hat{u}_{t-l} + \text{residual},$$

where the \hat{u}_{t-l} which cannot be computed are replaced by zero. The test statistic is the ordinary F statistic for all the ρ_l to be zero. It can be written

as

$$\frac{n - k - m - r}{r} \times \frac{\|\mathbf{P}_{\mathbf{M}_{[\mathbf{X} \ \mathbf{Y}]}\hat{\mathbf{V}}}\hat{\mathbf{u}}\|^2}{\|\mathbf{M}_{\mathbf{M}_{[\mathbf{X} \ \mathbf{Y}]}\hat{\mathbf{V}}}\hat{\mathbf{u}}\|^2}, \quad (28)$$

where $\hat{\mathbf{u}}$ has typical element \hat{u}_t , $\hat{\mathbf{V}}$ has typical element \hat{u}_{t-l} , $\mathbf{M}_{[\mathbf{X} \ \mathbf{Y}]}$ denotes the matrix that projects orthogonally off the space spanned by \mathbf{X} and \mathbf{Y} jointly, and $\mathbf{P}_{\mathbf{M}_{[\mathbf{X} \ \mathbf{Y}]}\hat{\mathbf{V}}}$ and $\mathbf{M}_{\mathbf{M}_{[\mathbf{X} \ \mathbf{Y}]}\hat{\mathbf{V}}}$ denote, respectively, the matrices that project on to and off the space spanned by $\mathbf{M}_{[\mathbf{X} \ \mathbf{Y}]}\hat{\mathbf{V}}$.

This approach is easy to implement, asymptotically valid, and asymptotically optimal against local alternatives. There is evidence that it works quite well in finite samples, somewhat better than asymptotically equivalent procedures that use χ^2 rather than F tests. However, the test statistic (28) is not exact in finite samples, and it is therefore natural to bootstrap it. The procedure is as follows:

1. Estimate (27) by OLS under the null hypothesis that $\rho_1 = \dots = \rho_r = 0$ so as to obtain $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\delta}}$, and $\hat{\mathbf{u}}$. Then construct $\hat{\mathbf{V}}$ from $\hat{\mathbf{u}}$ and compute the test statistic (28), which, following our earlier notation, we will call $\hat{\tau}$.
2. Draw B sets of bootstrap error terms, \mathbf{u}^* , and use them to generate B bootstrap samples \mathbf{y}^* . There are numerous ways in which the error terms can be drawn, four of which will be described below. The elements of \mathbf{y}^* are generated recursively from the equation

$$y_t^* = \mathbf{X}_t\hat{\boldsymbol{\beta}} + \mathbf{Y}_t^*\hat{\boldsymbol{\delta}} + u_t^*, \quad (29)$$

where the elements of \mathbf{Y}_t^* are equal to the observed values of \mathbf{Y}_t if they correspond to values of y_t prior to period 1, and equal to the appropriate lagged values of y_t^* otherwise.

3. For each bootstrap sample, compute τ^* using (28) with \mathbf{y}^* and \mathbf{Y}^* instead of \mathbf{y} and \mathbf{Y} . Then compute the estimated bootstrap P value as

$$\hat{p}(\hat{\tau}) = \frac{1}{B} \sum_{j=1}^B I(\tau_j^* \geq \hat{\tau}). \quad (30)$$

We consider four different ways of generating the u_t^* . For the parametric bootstrap, which we will call b_0 , they are simply independent draws from the $N(0, s^2)$ distribution, where s is the OLS estimate of σ from the regression run in step 1, that is, the square root of $SSR/(n - k - m)$. For the simplest nonparametric bootstrap, which we will call b_1 , they are obtained by resampling with replacement from the vector of \hat{u}_t . A slightly more complicated form of nonparametric bootstrap, which we will call b_2 , generates the u_t^* by resampling with replacement from the vector with typical element

$$(n/(n - k - m))^{1/2} \hat{u}_t.$$

The first factor here is a degrees of freedom correction. For both b_1 and b_2 , it is assumed that there is a constant among the regressors. If there were not, the residuals would have to be recentred and the consequent loss of one degree of freedom would have to be corrected for. Finally, the most complicated variety of nonparametric bootstrap, which we will call b_3 , generates the u_t^* by resampling from the vector with typical element \tilde{u}_t constructed as follows. First, divide each element of \hat{u}_t by the square root of one minus the t^{th} diagonal element of $\mathbf{P}_{[\mathbf{X} \ \mathbf{Y}]}$. Then recentre the vector that results and rescale it so that it has variance s^2 . This type of procedure has been advocated by Weber (1984) for bootstrapping regression models. In principle, it should reproduce the distribution of the original error terms more accurately than either b_1 or b_2 .

When $\boldsymbol{\delta} = \mathbf{0}$, so that there are no lagged dependent variables, the parametric bootstrap test b_0 is exact. In this case, under the null hypothesis, the test statistic (28) can be written as

$$\frac{n - k - r}{r} \times \frac{\|\mathbf{P}_{\mathbf{M}_X \hat{\mathbf{V}}} \mathbf{M}_X \mathbf{u}\|^2}{\|\mathbf{M}_{\mathbf{M}_X \hat{\mathbf{V}}} \mathbf{M}_X \mathbf{u}\|^2}, \quad (31)$$

which depends only on the matrix \mathbf{X} and the vector \mathbf{u} ; recall that each column of $\hat{\mathbf{V}}$ is just $\mathbf{M}_X \mathbf{u}$ lagged some number of times. The only parameter that affects \mathbf{u} is σ , and (31), like all F statistics, is invariant to its value. Thus (31) is pivotal. Note that, in this case, step 2 can be simplified, since (29) is no longer needed; we can just set the y_t^* equal to the u_t^* .

We have just shown that, for fixed regressors and normal errors, the parametric bootstrap test b_0 for serial correlation of any order is exact. This result is quite obvious, but it is also important. It provides a conceptually easy way to obtain valid, finite-sample P values for tests that applied econometricians use very frequently. Moreover, contrary to what some might expect, with modern computing technology this procedure is not at all computationally demanding. On a Pentium 90 personal computer (already somewhat out of date for people who take their computing seriously), it takes only 1.1 seconds for a reasonably efficient Fortran program to compute a test for AR(1) errors and its bootstrap P value for a model with 100 observations and 10 fixed regressors, using 999 bootstrap samples. If one of the regressors is a lagged dependent variable, the time rises somewhat, but only to 2.0 seconds.

The nonparametric bootstrap tests, b_1 through b_3 , will not be exact in the normal errors, fixed regressor case, but they will be asymptotically valid without the normality assumption. None of the tests will be exact when there are lagged dependent variables, since (28) does implicitly depend on all the parameters through the process that generates y_t recursively. However, our theoretical results suggest that all the tests should work very well. We now provide some evidence, based on Monte Carlo experiments, that provides strong support for this proposition.

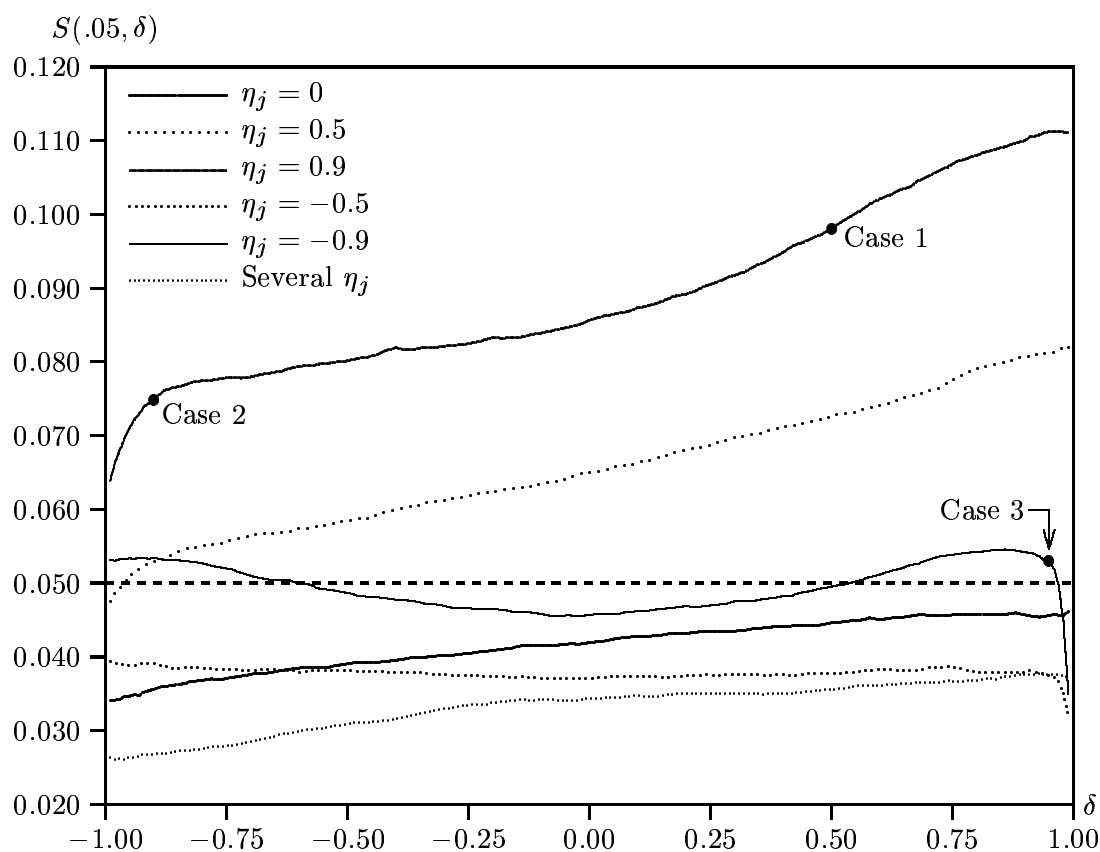


Figure 5. P Value Functions for AR(1) Tests at .05 Level, $n = 25$

All of our experiments dealt with a test for AR(1) errors in the context of a model with a constant term, four other exogenous variables, and a single lagged dependent variable. The four exogenous variables were generated from independent AR(1) processes with parameters η_j , $j = 1, \dots, 4$. We focused on the coefficient δ of the lagged dependent variable, setting all the β_i and σ to unity, because, if there were no lagged dependent variable, none of the other parameters would matter.

Figure 5 shows PVFs, as a function of δ , for $n = 25$ and various different choices of the η_j , based on 100,000 replications for each value of δ . These PVFs are constructed using the $t(n-7)$ distribution, since that is what most applied workers would use. It is evident that the characteristics of the \mathbf{X} matrix have a very substantial effect on the finite-sample performance of the test. We observe fairly severe overrejection in some cases, notably when all the η_j are equal to 0.9 and δ is large, quite good performance in other cases, and substantial underrejection in still others. For the PVF marked “several η_j ,” the four values were $-0.9, -0.5, 0.5,$ and 0.9 . Interestingly, this PVF is in no way an average of the others.

Figure 5 was used to decide what cases to investigate in depth. Case 1 was chosen as reasonably typical, since it has a plausible value of δ , 0.5, and not a great deal of curvature. On the other hand, Cases 2 and 3 were deliberately chosen to be ones where bootstrap tests might encounter problems, because the PVFs display considerable curvature. The values of δ are not very plausible, however, -0.9 for Case 2 and 0.95 for Case 3. It should be clear that the fact that the t distribution works very well for Case 3 does not imply that the bootstrap test will work well in this case. We also considered a fourth case, in which the parameters were the same as in Case 1, but the error terms had the $t(5)$ distribution instead of $N(0,1)$.

We computed the test statistic (28) and four sets of bootstrap P values (b_0 through b_3) for 15 different sample sizes: 8, 9, 10, 11, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, and 50. Each experiment used 100,000 replications, and there were $B = 999$ bootstrap samples for each replication. These numbers may seem rather large, but they were chosen for good reasons. As we shall see, the bootstrap tests work extremely well. Thus, in order to detect any pattern in the results, it is necessary to use a very large number of replications. It was necessary to use a fairly large value for B in order to avoid power loss. (Power will be discussed in a moment.)

The key results of our experiments are presented in Figure 6. Each panel shows the proportion of replications with P values less than .05 for each of the four bootstrap tests, as a function of n . The standard errors of these proportions, as estimates of test size, are about 0.00069. Results for the asymptotic tests are not shown, because the vertical scale would have had to be greatly compressed. It is clear from the figures that all the bootstrap tests work very well, except perhaps for $n = 8$, when the tests have just one degree of freedom. In Case 1, all the tests work essentially perfectly for $n \geq 9$. In Case 2, there seems to be a very slight tendency to overreject for most sample sizes, which is somewhat more severe for b_1 than for the other tests. This tendency is also evident in Case 3, where the performance of b_1 is a good deal worse than that of the other tests. In view of the fact that Cases 2 and 3 were deliberately chosen so that the bootstrap might encounter difficulties, the performance of all the tests is remarkably good. For Case 4, where the error terms are not normal, there seems to be a slight tendency for all the tests to underreject. This is most noticeable for the parametric bootstrap test, b_0 , which of course is not appropriate in this case.

Although all the bootstrap tests perform remarkably well, these results suggest that b_1 should be avoided and that b_2 or b_3 are the procedures of choice. They perform equally well, just about the same as the parametric bootstrap test b_0 in the cases where the latter is appropriate, and slightly better than the latter in Case 4, where b_0 is not appropriate. Since there seems to be no cost to using b_2 or b_3 when b_0 is appropriate, there seems to be no real reason to use the latter.

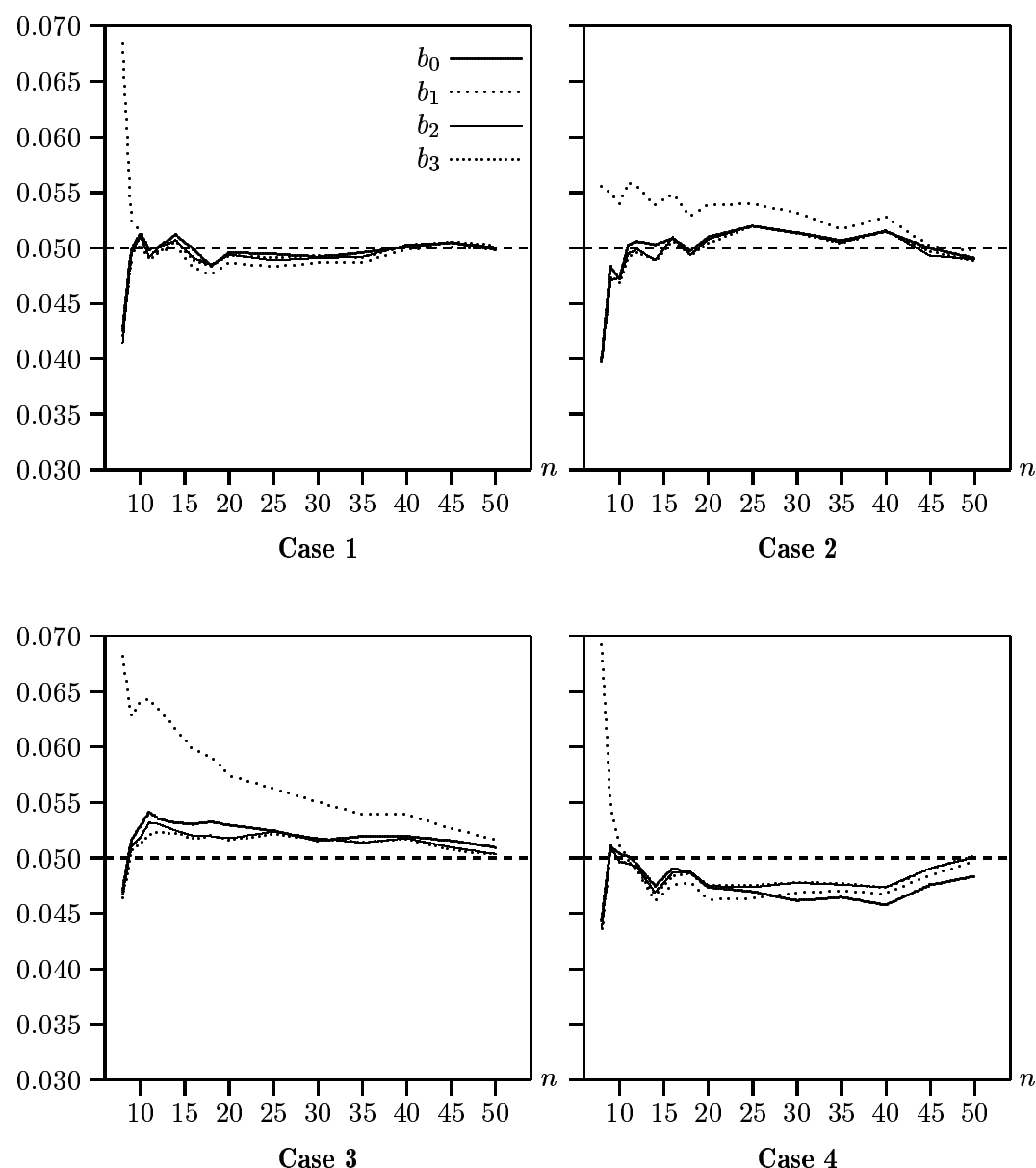


Figure 6. Estimated P Values for Bootstrap AR(1) Tests at .05 Level

We now turn our attention to power. It was mentioned earlier that, on a size-corrected basis, the power of the bootstrap test should be very similar to the power of the t test. However, there is no unique way to measure size-corrected power in a Monte Carlo experiment. All of our experiments involve a model with one lagged dependent variable and, possibly, AR(1) errors. Thus suppose that we generate experimental data using a DGP μ_1 with parameters β_1 , δ_1 , ρ_1 , and σ_1 . The results of this experiment will give us the nominal power of the test, but not the true power. To obtain the latter, we need to run

a matching experiment using a DGP that satisfies the null hypothesis. But what DGP should that be?

The obvious DGP to use is one with parameters β_1 , δ_1 , 0, and σ_1 . We shall call this DGP the **naive null**. There are at least two difficulties with the naive null. The first is that it may be a long way from the actual DGP, much farther than many other DGPs that also satisfy the null hypothesis. The second is that the naive null depends on the way the alternative model is parametrised. For instance, if instead of δ and ρ we were to use $\delta + \rho$ and ρ as parameters, then the naive null, in the old parametrisation, would be the DGP with parameters β_1 , $\delta_1 + \rho_1$, 0, and σ_1 . It would clearly be preferable to choose a DGP that satisfies the null in a parametrisation-independent fashion. Thus it is not at all clear that the size of the test under the naive null is what we want to use to compute size-corrected power.

Asymptotically at least, the closest null to a given fixed DGP is the null DGP characterized by the **pseudo-true values**, in the sense of White (1982), that correspond to the fixed DGP. The vector of pseudo-true values is defined as the probability limit of the quasi-maximum likelihood estimator of the null hypothesis under the fixed DGP. White shows that the pseudo-true values are the parameters of the DGP in the null hypothesis that minimize the **Kullback-Leibler Information Criterion (KLIC)** with respect to the fixed DGP. In practice, it is convenient simply to define the closest DGP in the null to be the one that minimizes the KLIC. In most cases of interest, although the KLIC formally depends on sample size, it turns out that the parameters of the KLIC-minimizing DGP are independent of the sample size. Note that the KLIC is a quantity defined purely in terms of two DGPs, quite independently of how these DGPs may be parametrised.

If we start from a given DGP μ_1 for a given sample size n , the drifting DGP through μ_1 suitable for power analysis has an end point in the null, μ_0 , which minimizes the KLIC to it from μ_1 . We will define the end point μ_0 as the **pseudo-true null**. Since we cannot perform a size correction of a nonpivotal test without choosing a specific null DGP, it appears that the pseudo-true null μ_0 is the most reasonable one to choose. While this choice is inevitably somewhat arbitrary, it has the advantages of being defined in a parametrisation-independent manner and of introducing no unnecessary dependence on the sample size. Moreover, Horowitz (1997) shows that a bootstrap test is asymptotically equivalent to an exact test of a simple null hypothesis consisting of just one DGP, namely the pseudo-true null. At least for bootstrap tests, this is another indication that the pseudo-true null is the most appropriate DGP to use for size correction even in finite samples.

With regard to the model (27), it would be quite easy to obtain the pseudo-true null if there were no lagged dependent variable, but its presence complicates matters considerably. However, it can be shown that the parameters of the pseudo-true null for this case may be obtained as follows.

First, regress $L(1 - \delta_1 L)^{-1} \mathbf{X} \boldsymbol{\beta}_1$ on \mathbf{X} and define \mathbf{b}_2 as the vector of parameter estimates and S as $1/n$ times the sum of squares of the residuals from that regression. (L denotes the lag operator.) Then the pseudo-true value of δ is

$$\delta_2 = \delta_1 + \frac{\rho_1 \sigma_1^2 (1 - \delta_1^2)}{AS + \sigma_1^2 (1 + \rho_1 \delta_1)},$$

where A is defined by

$$A = (1 - \rho_1 \delta_1)(1 - \delta_1^2)(1 - \rho_1^2).$$

The pseudo-true value of σ is the square root of

$$\sigma_2^2 = \frac{\sigma_1^2}{AS + \sigma_1^2 (1 + \rho_1 \delta_1)} \left(\frac{\sigma_1^2}{1 - \rho_1 \delta_1} + \frac{AS}{1 - \rho_1^2} \right).$$

Lastly, the pseudo-true value of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 - (\delta_2 - \delta_1) \mathbf{b}_2.$$

The experiments for power were somewhat less extensive than the ones for size. We first plotted power functions for Cases 1, 2, and 3 for various sample sizes, in order to be able to choose parameter values which would give the tests true power between 0.4 and 0.6. We did this because differences between the powers of different tests are most apparent when power is neither very large nor very small. For each case, we then picked various combinations of ρ and n to investigate. For each such combination, we ran three matched experiments with 50,000 replications each, using the same random numbers. For one of the three experiments, the DGP was the naive null, for another it was the pseudo-true null, and for the third it was the alternative with $\rho \neq 0$. Table 1 reports some results for $n = 15$, $n = 25$, and $n = 50$. Because the power functions were quite asymmetrical, it was often impossible to find values of ρ large enough to give power as large as 0.4 for the smaller sample sizes.

In order to calculate true power, we estimated power as a function of size, using local polynomial regressions in the neighborhood of size .05, and then calculated the fitted values at the point .05. The only column in Table 1 that directly reports power is the third column, which is marked $P(t_p)$. This is the power of the t test, calculated relative to size based on the pseudo-true null. The next column shows the difference between the power of the t test based on the naive null, t_n , and the power of the t test based on the pseudo-true null, t_p . Columns 5 and 6 show the difference between the powers of the bootstrap test, based on the pseudo-true and naive nulls, respectively, and $P(t_p)$. Finally, column 7 shows the difference between the two measures of power for the bootstrap tests. The bootstrap test results here are always for

Table 1. Power of AR(1) Tests

n	ρ	$P(t_p)$	$P(t_n) - P(t_p)$	$P(b_p) - P(t_p)$	$P(b_n) - P(t_p)$	$P(b_n) - P(b_p)$
Case 1						
15	-0.75	0.5351	-0.0096	0.0019	0.0038	0.0019
25	-0.45	0.4842	-0.0105	-0.0037	-0.0047	-0.0010
50	-0.25	0.4457	-0.0011	-0.0017	-0.0013	0.0004
50	0.50	0.4270	0.0059	0.0041	0.0039	-0.0002
Case 2						
15	-0.75	0.5685	-0.0160	0.0033	0.0009	-0.0023
25	-0.50	0.5788	-0.0053	-0.0018	-0.0029	-0.0011
50	-0.30	0.5435	0.0026	-0.0028	-0.0016	0.0012
50	0.50	0.5088	-0.0042	-0.0083	-0.0064	0.0018
Case 3						
15	0.90	0.6011	-0.0933	-0.0276	-0.0480	-0.0203
25	0.45	0.5618	-0.0527	-0.0530	-0.0544	-0.0014
50	0.25	0.4204	0.0057	0.0040	0.0077	0.0038
50	-0.45	0.4176	-0.0057	-0.0054	-0.0089	-0.0035

b_0 , the parametric bootstrap. We did obtain some results for b_2 , but these were always virtually indistinguishable from the results reported here.

There are at least two interesting results in Table 1. The first is that the differences between the powers of the t and bootstrap tests are generally very small. The exceptions are for Case 3, which was deliberately chosen to be a very difficult one, for $n = 15$ and $n = 25$. As the theory predicts, these generally small differences go in both directions; there is no reason to expect bootstrap tests to be systematically more or less powerful than asymptotic tests. The relatively large differences for Case 3 with small sample sizes arise because of the strange shape of the PVF in this case; see Figure 5. When $\rho > 0$, the pseudo-true value of δ is larger than its value in the naive null, and the critical values for the t test are quite different for these two nulls. Note that the difference between $P(t_n)$ and $P(t_p)$ is often greater than the difference between $P(b_p)$ and $P(t_p)$. In other words, how we measure true power makes a greater difference than whether or not we bootstrap.

The second interesting result in the table is that, with only one exception, the difference between $P(b_p)$ and $P(b_n)$ is always extremely small (less than .0040). This is precisely what the theory would lead us to expect. Because the bootstrap test does an excellent job of controlling size, it is close to being pivotal, and thus it does not matter very much whether we use the

naive or the pseudo-true null. The only exception is for Case 3 with $n = 15$, a deliberately extreme case.

The Monte Carlo results of this section strongly suggest that bootstrap tests for serial correlation work very well even when there are lagged dependent variables. As we noted above, for models with normal errors and without lagged dependent variables, the parametric bootstrap test b_0 works perfectly. The Monte Carlo results suggest that the nonparametric bootstrap tests b_2 and b_3 should work almost equally well.

Bootstrap J Tests

There are numerous procedures for testing nonnested regression models; for an introduction to the literature, see Davidson and MacKinnon (1993, Chapter 11). One of the simplest and most widely used is the J test proposed in Davidson and MacKinnon (1981). Like most nonnested hypothesis tests, this test is not exact in finite samples. Indeed, its finite-sample distribution can be very far from its asymptotic one. It therefore seems natural to bootstrap the J test.

For simplicity, we consider only the case of nonnested, linear regression models with IID normal errors. Suppose the two models are

$$\begin{aligned} H_1: \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_1, & \mathbf{u}_1 &\sim N(\mathbf{0}, \sigma_1^2 \mathbf{I}), \text{ and} \\ H_2: \mathbf{y} &= \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}_2, & \mathbf{u}_2 &\sim N(\mathbf{0}, \sigma_2^2 \mathbf{I}), \end{aligned}$$

where \mathbf{y} , \mathbf{u}_1 , and \mathbf{u}_2 are $n \times 1$, \mathbf{X} and \mathbf{Z} are $n \times k_1$ and $n \times k_2$, respectively, $\boldsymbol{\beta}$ is $k_1 \times 1$, and $\boldsymbol{\gamma}$ is $k_2 \times 1$. The J test statistic is the ordinary t statistic for $\alpha = 0$ in the artificial regression

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \alpha \mathbf{P}_Z \mathbf{y} + \text{residuals}, \quad (32)$$

where $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$. Thus $\mathbf{P}_Z \mathbf{y}$ is the vector of fitted values from least squares estimation of the H_2 model.

To bootstrap the J test, we first calculate the test statistic $\hat{\tau}$ by running regression (32) after obtaining the fitted values from the H_2 model. Then we use the parameter estimates from H_1 to generate B bootstrap samples. Using each of these bootstrap samples, we calculate a test statistic τ^* , and we then compute the estimated bootstrap P value via equation (30). As before, there are several ways in which the bootstrap samples can be generated. In our experiments, we used the parametric bootstrap b_0 and the three nonparametric bootstraps b_1 , b_2 , and b_3 , all of which were discussed above.

If \hat{s} denotes the estimated standard error from regression (32), the J test statistic can be written as

$$\frac{\mathbf{y}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{y}}{\hat{s}(\mathbf{y}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{y})^{1/2}}, \quad (33)$$

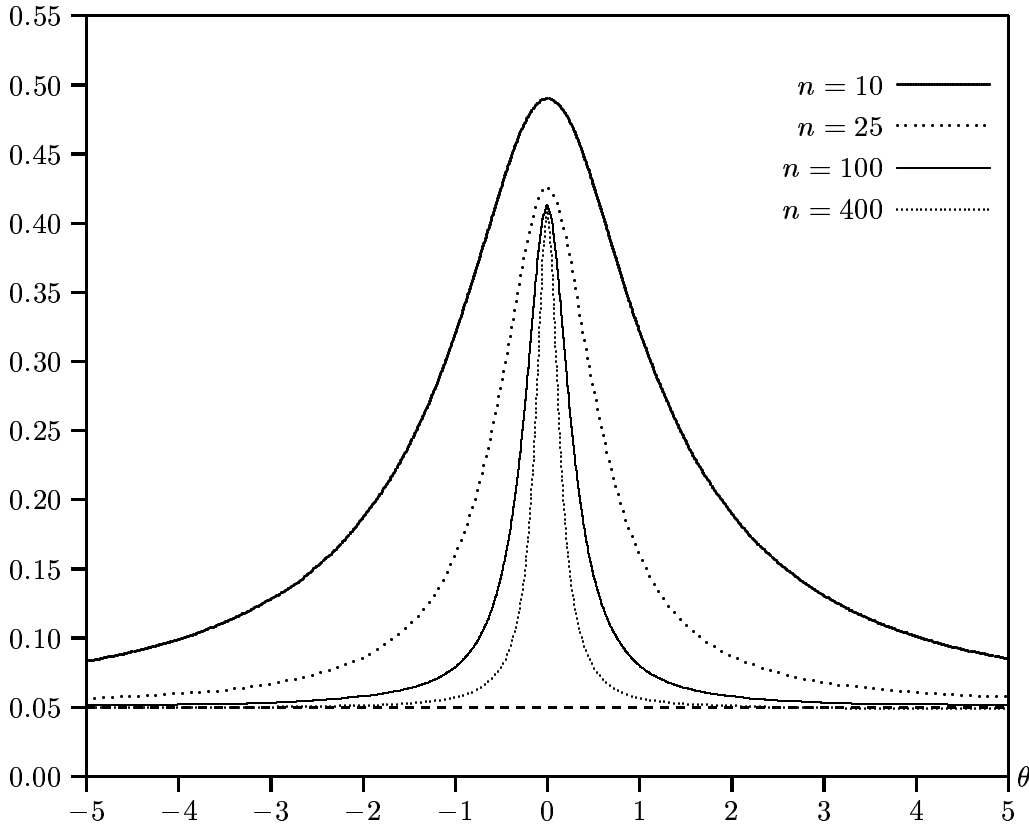


Figure 7. P Value Functions for J Tests

where $M_{\mathbf{X}} \equiv \mathbf{I} - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}$. It is straightforward to show that, under H_1 , the statistic (33) depends on both $\boldsymbol{\beta}$ and σ_1 , but only through the ratio $\boldsymbol{\beta}/\sigma_1$. Thus, if we choose a fixed vector $\bar{\boldsymbol{\beta}}$ and let $\boldsymbol{\beta} = \delta\bar{\boldsymbol{\beta}}$, the statistic will depend on a single parameter $\theta \equiv \delta/\sigma_1$. As we shall see in a moment, the finite-sample behavior of the test depends strongly on θ .

Our experiments were not intended to provide a comprehensive examination of the performance of the bootstrap J test. Such an examination is provided in Davidson and MacKinnon (1997). Instead, we deliberately chose a case for which the ordinary J test works badly, at least for some values of θ . We chose a simple scheme for generating \mathbf{X} and \mathbf{Z} . Each of the columns of \mathbf{X} , except for the constant term, was made up of IID normal random variables, was independent of the other columns, and was normalized to have length n . Each column of \mathbf{Z} was correlated with one of the columns of \mathbf{X} , with squared correlation 0.5 in most of our experiments. All elements of $\bar{\boldsymbol{\beta}}$ were equal.

Figure 7 shows P value functions for various values of n when $k_1 = 3$ and $k_2 = 6$. These are based on the t distribution with $n - 4$ degrees of freedom. The J test works relatively badly in this case, because there are 5 variables in \mathbf{Z} that are not in \mathbf{X} . For the smaller sample sizes, the performance of the

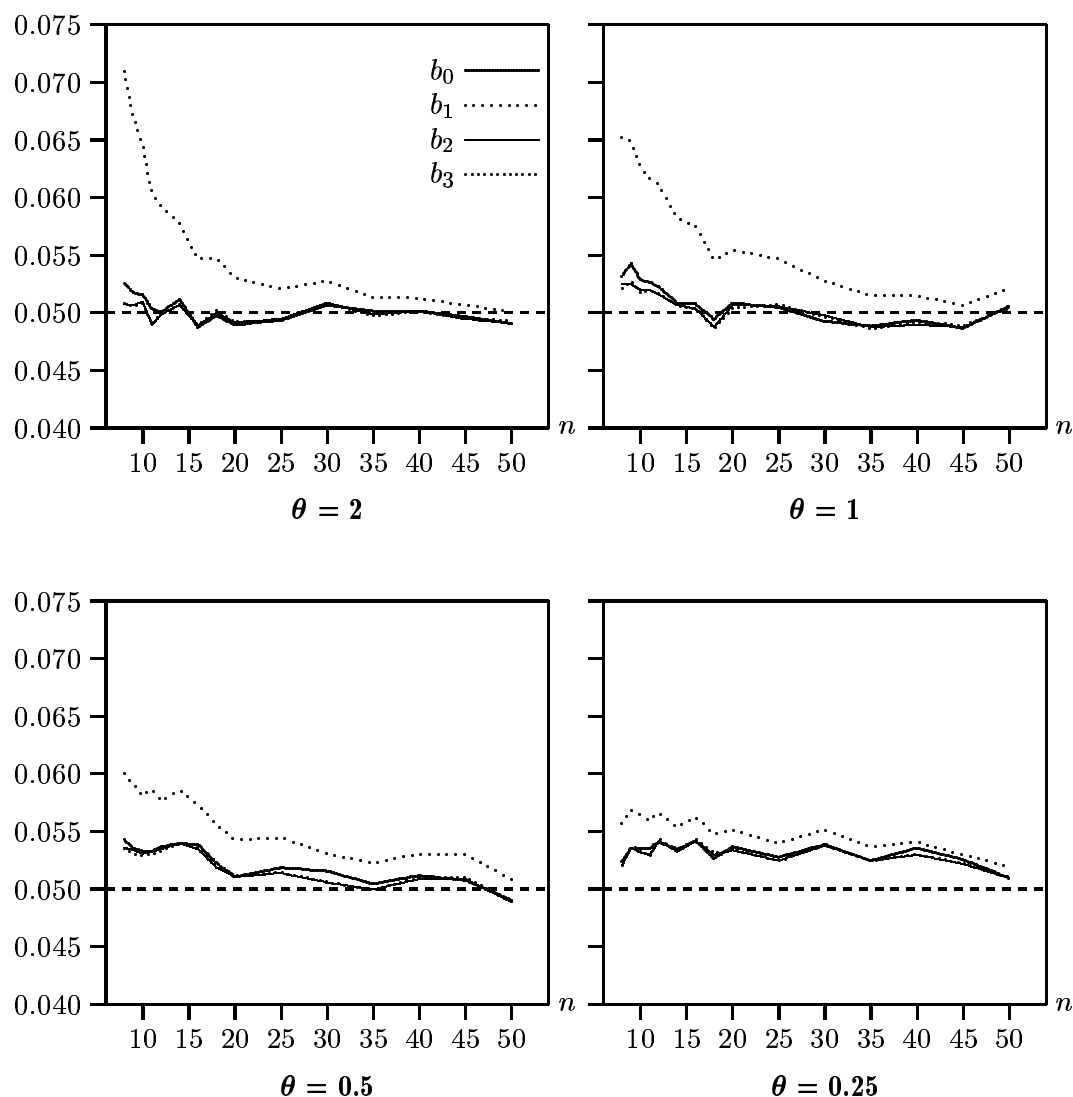


Figure 8. Estimated P Values for Bootstrap J Tests at .05 Level

J test is rather poor, except for quite large values of θ . For the larger sample sizes, the test generally performs much better, except near $\theta = 0$, where there is clearly a singularity. The usual asymptotic theory for the J test does not hold at this point, and we should not expect the usual theory of the bootstrap to hold either.

On the basis of Figure 7, one might reasonably expect that the bootstrap J test would work rather badly, because the PVF is very steep in many places and quite sharply curved in others. The results in Figure 8 may therefore come as a bit of a surprise. This figure shows the proportion of replications with P values less than .05, as a function of n , for the same sample sizes as before. Once again, the figure is based on 100,000 replications with $B = 999$.

For $\theta = 2$, all the tests except b_1 work essentially perfectly for $n \geq 10$. The reason b_1 works less well is that it implicitly uses an estimate of σ_1 that is biased downwards or, equivalently, an estimate of θ that is biased away from zero. It is easy to see from Figure 8 that this will cause the b_1 test to overreject. For $\theta = 1$, b_1 continues to perform poorly, but not quite as poorly, and the other tests continue to perform well, but not quite as well. They overreject slightly for very small values of n . For $\theta = 0.5$, b_1 performs a bit better, and the other tests perform less well, although still better than b_1 . The improvement of b_1 probably occurs because, as θ gets closer to zero, the PVF gets less steep, so the effect of bias diminishes. At the same time, the curvature increases, and this makes all the tests perform less well. Finally, for $\theta = 0.25$, which is quite close to the singularity, all the tests overreject for all values of n . Although this is very clear statistically, it is important to recognize that the extent of the overrejection is very modest indeed. For example, when $n = 25$, the b_0 and b_2 tests reject 5.27% and 5.25% of the time. In comparison, the t test rejects 37.90% of the time.

The bootstrap tests do not always work perfectly, but they do work extraordinarily well, and when they do not work perfectly the reason can usually be seen by looking at the PVF. These results by themselves do not show that bootstrap J tests will *always* work well. But the fuller results in Davidson and MacKinnon (1997) do show that, in almost all cases, they have very little size distortion.

5 Summary and Conclusion

The bootstrap provides higher-order refinements, relative to asymptotic theory, whenever the quantity bootstrapped is at least asymptotically pivotal. This is the case for all commonly used test statistics in econometrics. A refinement of order $n^{-1/2}$ is obtained whenever one computes the size distortion of a test, of given nominal size, based on a bootstrap P value. A further refinement, which in most cases will also be of order $n^{-1/2}$, is obtained whenever the test statistic is asymptotically independent of the bootstrap DGP, or, more specifically, of the appropriate quantile of the bootstrap distribution of the test statistic. Since most test statistics are indeed asymptotically independent of the estimates of the parameters of the null hypothesis produced by a wide class of extremum estimators, such test statistics, when bootstrapped, will benefit from this further degree of refinement. Thus bootstrap tests will, in many circumstances, be more accurate than asymptotic tests by a full order of n^{-1} .

Some of our more detailed results apply, strictly speaking, only to the case of the parametric bootstrap applied to a fully specified model. However, even nonparametric bootstrap DGPs usually depend on estimated parameters, and

they apparently give results indistinguishable from those of the parametric bootstrap in some circumstances, as with the two examples studied in detail above. Thus the analysis of the determinants of size and power of tests based on the parametric bootstrap is of general utility for judging when a bootstrap test is likely to behave badly.

A theoretical justification of this comes from Hall's (1992) Edgeworth expansion theory. There he shows, that, under suitable regularity conditions, the finite-sample distributions of test statistics, when considered up to the order of some power of $n^{-1/2}$, are determined fully in terms of just a few lower order moments of the statistic. In most cases, a nonparametric bootstrap DGP implicitly replaces these moments by sample moments from the original data. The relevant moments can be considered as parameters, and the sample moments are root- n consistent estimators of them. Thus the theory of the parametric bootstrap applies, to a high degree of approximation, to the nonparametric bootstrap also. An example of this is provided by the b_0 bootstrap, which implicitly uses a biased estimator of the error variance.

The P value discrepancy function is central to the results presented here. For given nominal size, this function measures, as a function of the actual DGP, the extent to which the actual level differs from the nominal level. Our principal results can be summarized, and understood intuitively, in terms of the properties of this function. The key point is that the probability that a bootstrap test will reject the null hypothesis for given nominal level α , whatever the actual DGP, is the probability, under that DGP, of a certain region in the space of the test statistic τ and the estimates of the model parameters θ . This region, which can be characterized purely in terms of the P value discrepancy function, is that in which the value of τ is greater than the level- α critical value of the DGP characterized by θ . It is thus just the region on one side of a level surface of the function.

For a given DGP satisfying the null hypothesis, the value of the P value discrepancy function is the size distortion of the asymptotic test. However, this value has no impact on the size of the corresponding bootstrap test. This is clear for pivotal statistics, for which the P value discrepancy function is constant, and the bootstrap test is exact. Even the first derivatives of the function, or equivalently the slope of its level surface, influence the size distortion of a bootstrap test, to leading order, only if the estimates of the parameters of the null hypothesis are biased. If they are not, then values of the estimates that would cause the bootstrap to overreject are compensated, to leading order, by values which would cause it to underreject. A bias in the parameter estimates would, however, cause one effect to dominate the other, and thus lead to a size distortion. With unbiased parameter estimates, the leading-order size distortion is determined by the second derivatives of the P value discrepancy function, that is, by the curvature of its level surface. Such curvature will once again cause values of the parameter estimates leading

to overrejection to have a greater or smaller impact than those leading to underrejection.

We also discussed what determines the size-corrected power of bootstrap tests compared with that of asymptotic tests. Size correction is much more of an issue for the latter than for the former, of course. We have seen that, once both tests are corrected for size, a bootstrap test can have different power from the corresponding asymptotic test only to the same order as the size distortion of the bootstrap test.

It is important to stress the fact that, although the size distortions of bootstrap tests that we have studied are real, they are remarkably small compared with those of asymptotic tests. In our Monte Carlo study, we went out of our way to seek situations in which the bootstrap would be ill-behaved. Even so, it was necessary to perform experiments of more than the usual accuracy, for very small sample sizes, in order to discern any evidence of misbehaviour, so as to provide confirmation of the theoretical results.

It is also important to stress the fact that, for many of the tests econometricians routinely use, the bootstrap is not, with modern computing technology, a very time-consuming procedure. We would urge the developers of econometric software to make the computation of bootstrap P values for such tests a standard feature of their programs, so that use of bootstrap tests might become routine.

References

- Barnard, G. A. (1963). "Contribution to discussion," *Journal of the Royal Statistical Society, Series B*, 25, 294.
- Beran, R. (1988). "Prepivoting test statistics: a bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, 83, 687–697.
- Davidson, R. and J. G. MacKinnon (1981). "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, 49, 781–793.
- Davidson, R. and J. G. MacKinnon (1987). "Implicit alternatives and the local power of test statistics," *Econometrica*, 55, 1305–29.
- Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, Oxford University Press.
- Davidson, R. and J. G. MacKinnon (1996a). "The size distortion of bootstrap tests," GREQAM Document de Travail No. 96A15.
- Davidson, R. and J. G. MacKinnon (1996b). "The power of bootstrap tests," Queen's University Institute for Economic Research, Discussion Paper 937.
- Davidson, R. and J. G. MacKinnon (1997). "Bootstrap Tests of Nonnested Linear Regression Models," Queen's University Institute for Economic Research, Discussion Paper 954.
- Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses," *Annals of Mathematical Statistics*, 28, 181–187.
- Efron, B. (1979). "Bootstrapping methods: another look at the jackknife," *Annals of Statistics*, 7, 1–26.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York, Springer-Verlag.
- Horowitz, J. L. (1997). "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in David M. Kreps and Kenneth F. Wallis, eds., *Advances in Economics and Econometrics: Theory and Applications*, Vol. 3, pp. 188–222, Cambridge, Cambridge University Press.
- Weber, N. C. (1984). "On resampling techniques for regression models," *Statistics and Probability Letters*, 2, 275–278.
- White, H. (1982). "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–26.

