Davidson, Russell; MacKinnon, James G.

Working Paper

The Size and Power of Bootstrap Tests

Queen's Economics Department Working Paper, No. 932

Provided in Cooperation with:
Queen’s University, Department of Economics (QED)

Suggested Citation: Davidson, Russell; MacKinnon, James G. (1996) : The Size and Power of Bootstrap Tests, Queen's Economics Department Working Paper, No. 932, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/189251

Queen's Economics Department Working Paper No. 932

The Size and Power of Bootstrap Tests

Russell Davidson James G. MacKinnon

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

2-1996

The Size and Power of Bootstrap Tests

by

Russell Davidson

and

James G. MacKinnon

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

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. This paper was written while the second author was visiting GREQAM.

February, 1996

Abstract

Bootstrap tests are tests for which the significance level is calculated by some sort of bootstrap procedure, which may be parametric or non-parametric. We show that, in many circumstances, the size distortion of a bootstrap P value for a test will be one whole order of magnitude smaller than that of the corresponding asymptotic P value. We also show that, at least in the parametric case, the magnitude of the distortion will depend on the shape of what we call the P value function. As regards the power of bootstrap tests, we show that the size-corrected power of a bootstrap test differs from that of the corresponding asymptotic test only by an amount of the same order of magnitude as the size distortion, and of arbitrary sign. Monte Carlo results are presented for two cases of interest: tests for serial correlation and nonnested hypothesis tests. These results confirm and illustrate the utility of our theoretical results, and they also suggest that bootstrap tests will often work extremely well in practice.

JEL Classification Number: C1

Keywords: bootstrapping, hypothesis testing, nonnested hypothesis tests, P values, tests for serial correlation

1. Introduction

Testing hypotheses is a central concern of classical econometrics. Sometimes the hypotheses to be tested are suggested by economic theory, and sometimes they are merely auxiliary hypotheses, such as homoskedasticity or serial independence, that must hold for inferences to be valid. Whichever the case, we want the tests to have the correct size and to have high power. Unfortunately, in the vast majority of interesting cases, the distributions of the test statistics we use are known only asymptotically. As a result, making inferences on the basis of them can be a risky undertaking.

There are two approaches to solving this problem. From a theoretical point of view, perhaps the most appealing is either to modify a test statistic analytically so that it approaches its asymptotic distribution more rapidly, as in Attfield (1995), or to modify the critical values so that the true size of the test approaches its nominal value more rapidly, as in Rothenberg (1984). Unfortunately, this approach often requires algebraic derivations that are very far from trivial, and in many cases it seems to be infeasible.

An alternative approach that is starting to become popular, largely because of the dramatic increase in the speeds of computers in recent years, is to employ some variant of the bootstrap. Although the statistical literature on bootstrapping is large and growing rapidly, most of it concerns confidence intervals rather than test statistics; see, among many others, Efron and Tibshirani (1993) and Hall (1992). The basic idea of bootstrapping a test statistic is to draw a large number of "bootstrap samples," which obey the null hypothesis and, as far as possible, resemble the real sample, and then compare the observed test statistic to the ones calculated from the bootstrap samples. An important recent paper which advocates this approach is Horowitz (1994).

In this paper, we prove a number of important results about the properties of what we shall call "bootstrap tests," that is, tests for which the significance level is calculated by some sort of bootstrap procedure. In the next section, we explain our terminology and notation and introduce several important concepts. The principal results of the paper are then proved in Sections 3, 4, and 5. Monte Carlo results which confirm and illustrate the utility of our theoretical results are presented in Sections 6 and 7. These deal, respectively, with tests for serial correlation and with the J test for nonnested hypotheses.

2. Some Basic Concepts

Suppose that we calculate a test statistic τ from a sample of size n. The details of how τ is calculated need not concern us, but it is essential that τ be asymptotically pivotal. In other words, the asymptotic distribution of τ under the null hypothesis must not depend on any unknown parameters. This is a rather weak assumption. All of the classical test statistics based on least squares, maximum likelihood, GMM, or other forms of extremum estimation satisfy an even stronger condition, since they actually have known asymptotic distributions.

It is possible to use bootstrapping either to calculate a critical value for τ or to calculate the significance level, or P value, associated with $\hat{\tau}$, the realized value of τ. We prefer the latter approach, partly 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, and partly because this approach leads naturally to the analysis of the next three sections. Our objective, then, is to compute

$$p(\hat{\tau}) \equiv \Pr_{\mu_0}(\tau \geq \hat{\tau}), \tag{1}$$

where μ_0 denotes the data generating process (DGP) which, under the null hypothesis, generated the data from which τ is calculated. Clearly, this DGP must itself satisfy the null hypothesis. In general, the probability in (1) depends on the sample size n and on the DGP μ_0. Only if $p(\hat{\tau})$ depends on neither of these will asymptotic theory give the right answer.

We may use either a parametric or a nonparametric bootstrap to draw the bootstrap samples. In the former case, we generate them from the model itself, using a vector of parameter estimates under the null, say $\hat{\boldsymbol{\theta}}$. This approach is appropriate in the case of a fully specified model, which we must have if we are using the method of maximum likelihood. In the latter case, in order to avoid imposing overly strong distributional assumptions, we resample from something like the empirical distribution function of the data. This approach is appropriate if the model is not fully specified, as in the case of GMM estimation. Actually, the term "nonparametric" may be somewhat misleading since, as we shall see in Sections 6 and 7, parameter estimates are often required to implement nonparametric bootstrap procedures. In either case, the DGP used to generate the bootstrap samples will depend on the sample used to obtain $\hat{\tau}$. We shall refer to this DGP as $\hat{\mu}$, the bootstrap DGP. In contrast, the actual (but unknown) DGP will be denoted μ_0.

Suppose we generate B bootstrap samples, each of size n, and use them to compute B test statistics, $\tau_j, j = 1, \ldots, B$. The bootstrap P value is defined as

$$p^*(\hat{\tau}) \equiv \Pr_{\hat{\mu}}(\tau \geq \hat{\tau}). \tag{2}$$

The only difference between (2) and (1) is that the former uses the bootstrap DGP $\hat{\mu}$ instead of the actual DGP μ_0 to compute the probability. By a law of large numbers,

$$p^*(\hat{\tau}) = \operatorname*{plim}_{B\to\infty} \frac{1}{B}\sum_{j=1}^{B} I(\tau_j > \hat{\tau}), \tag{3}$$

where $I(\cdot)$ is an indicator function, equal to 1 if its argument is true and equal to zero otherwise. Thus the bootstrap P value p^* is equal to the limit, as $B \to \infty$, of the proportion of bootstrap samples for which τ_j exceeds the observed test statistic $\hat{\tau}$.

In practice, of course, B will have to be finite, and we may wish to estimate p^* in a more efficient fashion than simply by using the finite-sample analogue of (3); see Davidson and MacKinnon (1996). Since the problems associated with finite B are unrelated to the issues discussed in this paper, and since, in many cases, it is feasible to make B so large that random errors in the bootstrapping process can safely be ignored, we shall in this paper be concerned solely with bootstrap P values based on an infinite number of bootstrap samples. Our objective is to understand the relationship between $p^*(\hat{\tau})$ and $p(\hat{\tau})$, both when the null hypothesis is true and when it is not.

One fundamental, and well-known, property of $p^*(\hat{\tau})$ is that, in the case of a parametric bootstrap, it is equal to $p(\hat{\tau})$ when τ is pivotal, provided of course that the parametric model is correct. In this case, the only difference between the DGP used to generate the τ_j and the DGP that is assumed to have generated $\hat{\tau}$ is that the former uses $\hat{\boldsymbol{\theta}}$ and the latter uses the true parameter vector $\boldsymbol{\theta}_0$. But if τ is pivotal, its distribution is the same for all admissible values of $\boldsymbol{\theta}$, and thus the same for both μ_0 and $\hat{\mu}$. Therefore, in this special case, $p^*(\hat{\tau}) = p(\hat{\tau})$.

Although this case is rather special, it has numerous applications in econometrics. For example, in univariate linear regression models with regressors that can be treated as fixed, any specification test that depends only on the residuals and the regressors will be pivotal. This includes many tests for serial correlation, heteroskedasticity, skewness, and kurtosis, including the information matrix test. Tests for serial correlation are discussed in Section 6. For the others, see Davidson and MacKinnon (1993, Chapter 16). Thus, provided the normality assumption is maintained, all of these commonly used tests can be made exact by using the parametric bootstrap.

The special case in which τ is pivotal makes it clear that it is only the fact that $\hat{\mu}$ differs from μ_0 which could cause bootstrap P values to be inaccurate. In order to understand the properties of bootstrap P values,

we need to describe how the distribution of τ depends on μ. We therefore define the *critical value function*, or CVF, $C(\alpha, \mu)$, by the equation

$$\Pr_\mu\big(\tau \geq C(\alpha, \mu)\big) = \alpha. \tag{4}$$

$C(\alpha, \mu)$ is thus the true level-α critical value for a test based on the statistic τ if the DGP is μ. In other words, it is the $1-\alpha$ quantile of the distribution of τ under μ. The rejection region for the bootstrap test is defined by

$$\hat{\tau} > C(\alpha, \hat{\mu}), \tag{5}$$

and the true size of the bootstrap test is simply the probability, under the DGP μ_0, of the event (5). This probability clearly depends only on the joint distribution under μ_0 of τ and the scalar quantity $C(\alpha, \hat{\mu})$. We shall make much use of this fact in the next three sections.

In the parametric case, μ will be a parametric model characterized by parameter vector $\boldsymbol{\theta}$, and we can, at least in principle, graph the CVF as a function of $\boldsymbol{\theta}$. As an illustration, Figure 1 shows the CVF for a particular test (a J test; see Section 7) with DGP characterized by a single parameter θ for $\alpha = .05$. The size of the bootstrap test is the probability mass, under θ_0, of a rejection region in the space of θ and τ. If $(\hat{\theta}, \hat{\tau})$ falls into this region, the bootstrap test rejects. From (5), it is clear that the rejection region consists of all values of $\hat{\tau}$ such that $\hat{\tau} > C(\alpha, \hat{\theta})$. It is the region above the graph of the CVF in (θ, τ) space.

The rectangle above the horizontal line marked $C(.05, 1)$ in the figure shows all $(\hat{\tau}, \hat{\theta})$ pairs that *should* lead to rejection at the .05 level when $\theta_0 = 1$. In contrast, the area above the CVF shows all pairs that actually *will* lead to rejection using a bootstrap test. How different these are will depend on the joint distribution of $\hat{\tau}$ and $\hat{\theta}$. For comparison, the rectangles above the two dotted lines show all pairs that will lead to rejection using the asymptotic critical value $C^\infty(.05) = 1.96$ and using the critical value $C^{22}(.05) = 2.074$ based on the $t(22)$ distribution. Clearly, the bootstrap test will work much better than either of these approximate tests.

From the figure, we see that when $\theta_0 = 1$, the bootstrap test will over-reject somewhat when $\hat{\theta} > 1$ and when $\hat{\theta} < -1$. For those values of θ, the CVF is below $C(.05, 1)$, and the bootstrap critical value will consequently be too small. By a similar argument, the bootstrap test will underreject when $-1 < \hat{\theta} < 1$. 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 $\theta = 1$. Thus, on average, we might expect the bootstrap test to work very well indeed in this case. This argument will be made much more precise in Section 5. In fact, as we shall see in Section 7, the bootstrap J test does work very well.

Because test statistics may have a wide variety of asymptotic distributions, and tests may be either one-tailed or two-tailed, it will be convenient for our analysis to convert any statistic into a corresponding approximate P value. Thus, instead of dealing with a statistic τ that has some known asymptotic distribution F, we shall deal with a statistic π_τ that will be asymptotically uniformly distributed on [0,1]. The asymptotic test would then reject at nominal size α if $\hat{\pi}_\tau < \alpha$. In the one-tailed case, we would have $\pi_\tau \equiv 1 - F(\tau)$.

In finite samples, of course, the event $(\pi_\tau < \alpha)$ will rarely have probability precisely α. Consequently, we introduce the P *value function*, or PVF, defined as follows:

$$S(\alpha, \mu) \equiv \Pr_\mu(\pi_\tau < \alpha). \tag{6}$$

For fixed μ, of course, $S(\alpha, \mu)$ is just the c.d.f. of π_τ evaluated at α. 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

$$S(\alpha, \mu) = \alpha + n^{-l/2} s(\alpha, \mu), \tag{7}$$

where the integer $l \geq 1$ is defined so that $s(\alpha, \mu)$ will be $O(1)$. In most cases, we expect that $l = 1$, but there may be exceptions. Note that $s(\alpha, \mu)$ will be independent of the DGP μ only if the statistics τ, and hence π_τ, are pivotal.

It is possible to compute a critical value function for π_τ rather than for τ. Since one usually rejects the null hypothesis when π_τ is too small, the analogue of (4) when dealing with an approximate P value is

$$\Pr_\mu\big(\pi_\tau < Q(\alpha, \mu)\big) = \alpha, \tag{8}$$

where Q denotes the CVF for π_τ. The CVF $C(\alpha, \mu)$ for τ and the CVF $Q(\alpha, \mu)$ for π_τ are related by

$$Q(\alpha, \mu) = 1 - F\big(C(\alpha, \mu)\big).$$

Both sides of this equation can be interpreted as the nominal size of a test that has true size α. They can also be interpreted as the α quantile of the random variable π_τ under the DGP μ.

The CVF $Q(\alpha, \mu)$ can be thought of as the inverse of the P value function $S(\alpha, \mu)$, in the sense that

$$S\big(Q(\alpha, \mu), \mu\big) = \alpha, \tag{9}$$

a result that follows immediately from evaluating (6) at $\alpha = Q(\alpha, \mu)$ and then using the definition (8) of $Q(\alpha, \mu)$. From (9) we also obtain

$$Q\big(S(\alpha, \mu), \mu\big) = \alpha,$$

since both S and Q are increasing in their first arguments. Analogously to (7), we will have

$$Q(\alpha, \mu) = \alpha + n^{-l/2} q(\alpha, \mu), \tag{10}$$

with the function q of order unity. The integer l will be the same as the l in (7). Figure 2 graphs the PVF $S(.05, \theta)$ and its inverse $Q(.05, \theta)$ for exactly the same one-parameter case as the CVF in Figure 1. All three functions evidently convey essentially the same information.

Many of the basic ideas of this paper can be understood from Figures 1 and 2. Whenever a test is not pivotal, the CVF for the test, and hence also the PVF and its inverse, will not be flat. As a consequence, a bootstrap test, because it is based on the bootstrap DGP $\hat{\mu}$, will almost never have exactly the right size. However, as we demonstrate in the next section, there are good reasons to believe that bootstrap tests will very often have almost the right size. It is clear from the figures that the size of a bootstrap test based on a test statistic τ must depend on the *joint* distribution of τ and $\hat{\mu}$. It is the need to deal analytically with this dependence that makes the analysis of the next three sections a little bit difficult.

Before we move on to the next section, we present a much simpler analysis. If we simply condition on $\hat{\mu}$, it is very easy, at least in the parametric case, to see that bootstrap tests will perform better than asymptotic tests. Suppose that the DGP μ is fully characterized by a parameter vector $\boldsymbol{\theta}$, so that the P value discrepancy function can be written as $n^{-l/2} s(\alpha, \boldsymbol{\theta})$. A first-order Taylor expansion of the latter around the true parameter vector $\boldsymbol{\theta}_0$ yields

$$n^{-l/2} s(\alpha, \hat{\boldsymbol{\theta}}) \stackrel{a}{=} n^{-l/2} s(\alpha, \boldsymbol{\theta}_0) + n^{-l/2} \boldsymbol{s}^\top(\alpha, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{11}$$

where $\boldsymbol{s}(\alpha, \boldsymbol{\theta}_0)$ is the vector of first derivatives of $s(\alpha, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}_0$. From (7), the difference between the probability that $\hat{\pi}_\tau < \alpha$ according to the bootstrap DGP and the true probability is

$$S(\alpha, \hat{\boldsymbol{\theta}}) - S(\alpha, \boldsymbol{\theta}_0) = n^{-l/2}\big(s(\alpha, \hat{\boldsymbol{\theta}}) - s(\alpha, \boldsymbol{\theta}_0)\big). \tag{12}$$

Substituting (11) into (12) yields

$$S(\alpha, \hat{\boldsymbol{\theta}}) - S(\alpha, \boldsymbol{\theta}_0) \stackrel{a}{=} n^{-l/2} \boldsymbol{s}^\top(\alpha, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{13}$$

If $\hat{\boldsymbol{\theta}}$ is root-n consistent, the quantity on the right-hand side of (13) is of order $n^{-(l+1)/2}$. Thus the bootstrap test appears to reduce the order of the P value discrepancy by a factor of $n^{-1/2}$.

The above argument is perfectly correct as far as it goes. The problem with it is that it involves conditioning on $\hat{\boldsymbol{\theta}}$, that is, $\hat{\mu}$. It therefore does not take into account the randomness of $\hat{\mu}$, an essential feature of bootstrap testing. As we shall see in the next section, when this randomness is taken into account, the bootstrap test is seen to perform even better than the simple analysis above suggests.

3. The Size of Bootstrap Tests

In this section, we obtain the order of the P value discrepancy function for bootstrap tests. For the nonparametric bootstrap, the answer is not entirely straightforward, since it depends on the properties of the bootstrap DGP $\hat{\mu}$ as an estimator of the actual DGP μ_0, and these will depend on the precise bootstrapping procedure used. For the parametric bootstrap, the answer is clearer. If it is based on parameter estimates under the null, the parametric bootstrap will be at least one full order more accurate than the asymptotic test. This type of result is not entirely without precedent in the statistical literature. However, our results are both much more detailed and more general than those of, for example, Hall and Titterington (1989) and Hall (1992).

The bootstrap critical value for π_τ, $Q(\alpha, \hat{\mu})$, is a random variable which will be asymptotically nonrandom and equal to α. In finite samples, its value should generally be close to $Q(\alpha, \mu_0)$ for two reasons. The first reason is that, if π_τ is nearly pivotal, then $Q(\alpha, \mu)$ does not depend much on μ. The second reason is that $\hat{\mu}$ will generally be close to μ_0. It is therefore convenient to define a new random variable γ, of order unity as $n \to \infty$, as follows:

$$Q(\alpha, \hat{\mu}) = Q(\alpha, \mu_0) + n^{-k/2}\sigma_\gamma \gamma, \tag{14}$$

where σ_γ is independent of n and is chosen so that γ has variance unity asymptotically, and where k is an integer chosen to make (14) true.

In the case of the parametric bootstrap based on root-n consistent estimates, $k = l + 1$. Recall from (13) that the difference between $S(\alpha, \hat{\mu})$ and $S(\alpha, \mu_0)$ in this case is $O(n^{-(l+1)/2})$. Since $Q(\alpha, \mu)$ is just the inverse of $S(\alpha, \mu)$, $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$ must also be $O(n^{-(l+1)/2})$. Thus, since $l \geq 1$, we can be confident that $k \geq 2$ for the parametric case. There is also good reason to believe that $k \geq 2$ in most nonparametric cases of interest. What is needed is that we should be able to write (10) not only for μ satisfying the null hypothesis, but also for the μ that are used as nonparametric

bootstrap distributions. Then, provided that $\hat{\mu} - \mu_0 = O(n^{-1/2})$, as will normally be the case for any sensible nonparametric bootstrap, it is clear that $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$ will be $O(n^{-(l+1)/2})$.

The P value function $S(\alpha, \mu)$, defined in (6), can be interpreted as the c.d.f. of π_τ under μ. In order to describe the *joint* distribution of π_τ and γ, we also need the distribution of γ conditional on π_τ. Let us denote by $g(\gamma, \pi_\tau)$ the density of γ conditional on π_τ under the DGP μ_0. Since $g(\gamma, \pi_\tau)$ is a density,

$$\int_{-\infty}^{\infty} g(\gamma, \pi_\tau)\, d\gamma = 1 \quad \text{for all } \pi_\tau \in [0, 1]. \tag{15}$$

With this specification, we can compute the true size of the bootstrap test as the probability under μ_0 that $\pi_\tau < Q(\alpha, \hat{\mu})$. The true size is

$$\int_{-\infty}^{\infty} d\gamma \int_0^{Q + n^{-k/2}\sigma_\gamma \gamma} dS(\pi_\tau)\, g(\gamma, \pi_\tau), \tag{16}$$

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

The integral over π_τ in (16) can be split into two parts, as follows:

$$\begin{aligned} &\int_0^{Q} dS(\pi_\tau) \int_{-\infty}^{\infty} d\gamma\, g(\gamma, \pi_\tau) \\ &\qquad + \int_{-\infty}^{\infty} d\gamma \int_0^{n^{-k/2}\sigma_\gamma \gamma} d\pi_\tau\, S'(Q + \pi_\tau)\, g(\gamma, Q + \pi_\tau), \end{aligned} \tag{17}$$

where S' is the derivative of $S(\pi_\tau)$. Because of (15), the integral over γ in the first term of (17) equals 1, and so the whole first term equals α, by (9). Since the nominal size of the bootstrap test is α, the size discrepancy for the test is given by the second term in (17). This second term can be written as

$$n^{-k/2}\sigma_\gamma \int_{-\infty}^{\infty} d\gamma\, \gamma\, g(\gamma, \alpha) + O(n^{-(k+1)/2}), \tag{18}$$

since, by (10), $Q \equiv Q(\alpha, \mu_0) = \alpha + O(n^{-l/2})$, and this, along with (7), gives $S'(Q) = 1 + O(n^{-l/2})$. Thus the true size of the bootstrap test is α plus the two terms in (18).

The first term in (18) has a simple interpretation. It is the expectation, conditional on $\pi_\tau = \alpha$, of $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$. Thus it is the bias, conditional on $\pi_\tau = \alpha$, of the size-α critical value based on the bootstrap DGP $\hat{\mu}$. When this bias is nonzero, it is responsible for the size distortion

of the bootstrap test to leading order. On the other hand, when this bias is zero, the first term in (18) vanishes, and the size distortion of the bootstrap test is of lower order than otherwise. In that case, those $\hat{\mu}$ which overestimate the size-α critical value will be balanced on average by those $\hat{\mu}$ which underestimate it. Actually, the bias does not have to be zero. Provided the bias is $O(n^{-(k+1)/2})$ or lower, the leading-order term in (18) will be $O(n^{-(k+1)/2})$. Unfortunately, the bias of $Q(\alpha, \hat{\mu})$ will generally be of order $O(n^{-k/2})$ if it is not zero.

We can obtain an even stronger result than (18) if a further condition is satisfied. This condition is that γ and π_τ should be asymptotically independent. For the parametric bootstrap, this condition will always be satisfied, provided the parameters are estimated under the null and have the usual asymptotic properties. To see this, observe that, in the parametric case, $Q(\alpha, \hat{\mu})$, and hence γ, is simply a function of the vector of parameter estimates under the null, which once again we may call $\hat{\boldsymbol{\theta}}$. Asymptotically, if $\hat{\boldsymbol{\theta}}$ is an extremum estimator that satisfies first-order conditions in the interior of the parameter space, the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ will be asymptotically independent of any classical test statistic. Hence γ must be asymptotically independent of π_τ.

We shall not attempt a detailed proof of the independence result here. Essentially, we need to express $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ as a linear function of the gradient vector and the test statistic as a quadratic function of that same vector; see, for example, expressions (13.19) and (13.23) of Davidson and MacKinnon (1993). We would then show that these two expressions can be written as $\boldsymbol{a}^\top \boldsymbol{u}$ and $\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}$, respectively, where $\boldsymbol{a}$ is a nonrandom vector, $\boldsymbol{A}$ is a nonrandom matrix, $\boldsymbol{u}$ is $N(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{a}^\top \boldsymbol{A} = \mathbf{0}$. The normality of $\boldsymbol{u}$ and the fact that $\boldsymbol{a}^\top \boldsymbol{A} = \mathbf{0}$ implies that $\boldsymbol{a}^\top \boldsymbol{u}$ and $\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}$ are independent, which implies that the vector of parameter estimates and the test statistic are asymptotically independent. This independence can be proved for ML, NLS, GMM, and other forms of extremum estimation. For the case of the classical test statistics based on maximum likelihood estimation, a detailed proof may be found in Davidson and MacKinnon (1987).

If we assume that γ and π_τ are asymptotically independent, then we can write

$$g(\gamma, \pi_\tau) = h(\gamma)\big(1 + n^{-j/2} f(\gamma, \pi_\tau)\big), \tag{19}$$

where $h(\gamma)$ is the asymptotic marginal distribution of γ, $j \geq 1$ is a suitable integer, and $f(\gamma, \pi_\tau)$ is of order unity as $n \to \infty$. In the most usual case, $j = 1$, although larger values of j are possible.

Since $\hat{\mu}$ must be a consistent estimator of μ_0, so that $Q(\alpha, \hat{\mu})$ must be a consistent estimator of $Q(\alpha, \mu_0)$, we have that

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

Thus (18) becomes

$$(20) \qquad n^{-(k+j)/2} \sigma_\gamma \int_{-\infty}^{\infty} d\gamma \, \gamma \, h(\gamma) \, f(\gamma, \alpha) + O(n^{-(k+j+1)/2}).$$

The interpretation of (20) is the same as that of (18). The first term is the bias, conditional on $\pi_\tau = \alpha$, of the size-α critical value based on the bootstrap DGP $\hat{\mu}$. In the usual case, with $k = 2$ and $j = 1$, this term will be $O(n^{-3/2})$.

Rather than attempting to state a formal theorem, let us summarize the results of this section. The key is the order of the discrepancy between the bootstrap critical value and the true critical value. In (14), we specified that it was $O(n^{-k/2})$. In the parametric case with root-n consistency, $k = l + 1$ with $l \geq 1$, the usual situation being $l = 1$ and $k = 2$. In the nonparametric case, the value of k is less clear, but any sensible procedure should have $k \geq 2$. In (18), we proved that, in the worst case, the bootstrap P value will be incorrect only at $O(n^{-k/2})$. This is also what we obtained by the simple analysis of the last section, which was only for the parametric case and was conditional on $\hat{\mu}$.

More interestingly, we showed that the errors in bootstrap P values will often be of smaller order than this. First of all, when the bootstrap critical value is unbiased to highest order, the bootstrap P value will be incorrect only at $O(n^{-(k+1)/2})$. Secondly, when the discrepancy is asymptotically independent of the test statistic π_τ, the bootstrap P value will be incorrect only at $O(n^{-(k+j)/2})$, where $j/2$ is the highest order at which dependence occurs. This second case always holds for the parametric bootstrap if it is based on a regular extremum estimator under the null, and it undoubtedly holds for many nonparametric bootstrap procedures as well. In the usual case, with $l = 1$, $k = 2$, and $j = 1$, we thus have two, potentially quite common, situations in which the error in the bootstrap P value will be $O(n^{-3/2})$ when the error in the asymptotic P value is $O(n^{-1/2})$.

4. The Power of Bootstrap Tests

In this section, we characterize the difference between the power of bootstrap tests and the power of the asymptotic tests on which they are based. The analysis is in many respects very similar to that of the previous section, except that the DGP is assumed to be μ_1, which is not a member of the null hypothesis. The test statistic in P value form, π_τ, will no longer have a distribution close to $U(0,1)$, at least not if the test has any reasonable power. In fact, if μ_1 is a fixed DGP, independent of the sample size, then π_τ will be asymptotically concentrated on zero, since any consistent test will asymptotically reject the null hypothesis with probability one. For asymptotic theory to give sensible results, it is therefore usual to postulate a *drifting DGP*, that is, one which is determined by a DGP belonging to the null hypothesis plus a perturbation of order $O(n^{-1/2})$; see Davidson and MacKinnon (1993, Chapter 12) for a detailed discussion of drifting DGPs.

For any given sample size, the c.d.f. of π_τ under μ_1 can be written as

$$P(\alpha, \mu_1) \equiv \Pr_{\mu_1}(\pi_\tau < \alpha). \tag{21}$$

This definition of $P(\alpha, \mu_1)$ is very similar to the definition of $S(\alpha, \mu)$ in (6). As before, we may sometimes drop the second argument. The notation has changed slightly because now we are concerned with power rather than size. For α different from 0 or 1, $P(\alpha, \mu_1)$ will tend neither to zero nor to infinity as $n \to \infty$, because μ_1 drifts towards the null hypothesis at an appropriate rate.

Unlike μ_1, the bootstrap distribution $\hat{\mu}$ must belong to the null hypothesis, or at least, in the nonparametric case, it must be close to it in the appropriate sense. It is a random distribution determined by μ_1 rather than by some μ_0 in the null hypothesis, but it is just the same sort of distribution as it would have been if it had been determined by such a μ_0. In particular, for a parametric bootstrap, it is determined by estimates of the parameters of the null hypothesis, or, for a nonparametric bootstrap, by residuals or similar quantities obtained by estimating the null hypothesis. If we consider the α quantile of $\hat{\mu}$, we can see that its general properties as a random variable are just as they were in the previous section. It will be asymptotically nonrandom and equal to α, and it will be expressible, as in (14), in terms of $Q(\alpha, \mu_0)$ and an asymptotically mean zero, variance unity, random variable that we can still write as γ.

It is not entirely clear just what μ_0 to use in equation (14) when the actual DGP is μ_1. Essentially, we want μ_0 to be as close as possible to μ_1 while still satisfying the null hypothesis. From the theoretical point of view, μ_0 must simply be such that $Q(\alpha, \hat{\mu})$ is given by (14). However, that equation is not quite enough to determine μ_0 uniquely, since changing μ_0 by

an amount that affects $Q(\alpha, \mu_0)$ only by a quantity of order $O(n^{-(k+1)/2})$ is clearly compatible with all the requirements on k and on γ. For the Monte Carlo work to be discussed in Section 6, the precise choice of μ_0 does matter, and this issue will be discussed further there.

Let us suppose then that, under the drifting DGP μ_1, equation (14) is satisfied for some μ_0 and for some γ that asymptotically has mean zero and variance unity. As before, we need the joint density of π_τ and γ under μ_1. By (21), the marginal density of π_τ is $P'(\pi_\tau)$, and we may denote the density of γ conditional on π_τ by $g_1(\gamma, \pi_\tau)$. The power of the bootstrap test based on π_τ at nominal size α is the probability under μ_1 of rejecting the null hypothesis, that is, the probability that $\pi_\tau < Q(\alpha, \hat{\mu})$. As in (16), the power can be expressed as

$$\int_{-\infty}^{\infty} d\gamma \int_0^{Q(\alpha,\mu_0)+n^{-k/2}\sigma_\gamma\gamma} dP(\pi_\tau)\, g_1(\gamma, \pi_\tau), \tag{22}$$

and, as in (17), this can be split into two parts. The first part is

$$\int_0^{Q} dP(\pi_\tau) \int_{-\infty}^{\infty} d\gamma\, g_1(\gamma, \pi_\tau) = P\big(Q(\alpha, \mu_0)\big). \tag{23}$$

This is simply the power of the asymptotic test based on π_τ, where the critical value is such that the test has true size α under μ_0. It can therefore be interpreted as the size-corrected power of the asymptotic test at level α.

The second term in (22) is

$$\int_{-\infty}^{\infty} d\gamma \int_0^{n^{-k/2}\sigma_\gamma\gamma} d\pi_\tau\, P'\big(Q(\alpha, \mu_0) + \pi_\tau\big)\, g_1\big(\gamma, Q(\alpha, \mu_0) + \pi_\tau\big). \tag{24}$$

To leading order, this is just

$$n^{-k/2}\sigma_\gamma\, P'(\alpha) \int_{-\infty}^{\infty} d\gamma\, \gamma\, g_1(\gamma, \alpha) + O(n^{-(k+1)/2}); \tag{25}$$

compare (18). This term is the difference in power between the bootstrap test at *nominal* level α and the asymptotic test at *true* level α.

What we are interested in is the difference in power between the bootstrap and asymptotic tests at true level α. For this, we must replace α in (23) and (24) by the nominal level which corresponds to true level α for the bootstrap test. The appropriate correction is given by (18), and so α is to be replaced by

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

To the order we are considering, this replacement affects only (23), which becomes to leading order

$$P\big(Q(\alpha,\mu_0)\big) - n^{-k/2}\sigma_\gamma\, P'(\alpha) \int_{-\infty}^{\infty} d\gamma\; \gamma\, g(\gamma,\alpha).$$

From this and from (25), we conclude that the difference between the power of the bootstrap and asymptotic tests at true level α is

$$(26) \qquad n^{-k/2}\sigma_\gamma\, P'(\alpha) \int_{-\infty}^{\infty} d\gamma\; \gamma\, \big(g_1(\gamma,\alpha) - g(\gamma,\alpha)\big) + O(n^{-(k+1)/2}).$$

The result, mentioned in the last section, that extremum estimators of parameter estimates under the null hypothesis are asymptotically independent of the classical test statistics holds more generally for any drifting DGP of the sort we consider in this section. This is shown in some generality for maximum likelihood estimators in Davidson and MacKinnon (1987). Consequently, just as $g(\gamma,\alpha)$ may be expressed as in (19), so may we write

$$g_1(\gamma,\alpha) = h_1(\gamma)\big(1 + n^{-j/2} f_1(\gamma,\alpha)\big),$$

where, as with h,

$$\int_{-\infty}^{\infty} \gamma\, h_1(\gamma)\, d\gamma = 0$$

because γ has asymptotic mean zero. The leading term in (26) will thus be of order $O(n^{-(k+j)/2})$; compare (20). Specifically, (26) becomes

$$(27) \quad n^{-(k+j)/2}\sigma_\gamma\, P'(\alpha) \int_{-\infty}^{\infty} d\gamma\; \gamma\, \big(f_1(\gamma,\alpha) - f(\gamma,\alpha)\big) + O(n^{-(k+j+1)/2}).$$

From expressions (26) and (27), we obtain two important and very simple results. First of all, we see that, in general, the bootstrap and asymptotic tests will have power that differs, on a size-corrected basis, only at $O(n^{-k/2})$. In the case of the parametric bootstrap, this means that any discrepancy is at most $O(n^{-(l+1)/2})$, a result similar to one obtained by Horowitz (1994). An even stronger result holds in cases involving classical test statistics, for which the parametric bootstrap is always available. In such cases, the discrepancy is the same order as the size discrepancy obtained in the last section, namely $O(n^{-(k+j)/2})$ or, for the usual case with $k = 2$ and $j = 1$, $O(n^{-3/2})$. Secondly, we see from the first terms in (26) and (27) that, to highest order, any power difference is due solely to the possibility that γ may have a different distribution under the null and under the nonnull DGPs. In other words, the power difference arises from the possibility that $Q(\alpha,\hat{\mu}) - Q(\alpha,\mu_0)$ may differ under μ_0 and μ_1. For a pivotal test, there can be no power difference, and for a test that is reasonably close to being pivotal, we would expect the difference to be very small.

5. The One-Parameter Case

In this section, we derive, to highest order, the P value discrepancy function for a parametric bootstrap test when the DGP μ depends only on a single parameter θ. This result turns out to be easily interpretable and very useful in understanding the behavior of bootstrap tests, even nonparametric ones.

We assume that $\hat{\theta}$ is a root-n consistent, asymptotically normal estimator of the parameter θ_0. In this simple case, the bootstrap distribution $\hat{\mu}$ is just the DGP characterized by $\hat{\theta}$. Since we are considering only one parameter under the null, we suppose that it has been subjected to a variance-stabilizing transformation such that, under θ_0,

$$e \equiv n^{1/2}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} N(0,1).$$

It will be convenient to work with the random variable e instead of $\hat{\theta}$. As we saw in Section 3, it is reasonable to assume that π_τ is asymptotically pivotal and that π_τ and e are asymptotically independent. Then the joint density of π_τ and e can be written as

$$S'(\pi_\tau, \theta_0)\phi(e)\big(1 + n^{-1/2}a(e, \pi_\tau)\big). \tag{28}$$

Here, as in Section 3, S' denotes the derivative of S with respect to its first argument, and as usual, $\phi(\cdot)$ denotes the density of the $N(0,1)$ distribution. We require that $a(e, \pi_\tau) = O(1)$. In principle, we could have $n^{-j/2}$ instead of $n^{-1/2}$ in (28), as we did in (19). However, it simplifies the result considerably to assume that $j = 1$.

What we wish to do now is to obtain explicit expressions for the factors that appear in the general result (20) for the P value discrepancy function of the bootstrap test, and then substitute them into that expression. Because the details of the derivation are somewhat tedious, they are relegated to the Appendix. The final result is quite simple, however. The P value discrepancy for the bootstrap test at size α is

$$-n^{-(l+2)/2}\left(s_\theta(\alpha,\theta)\int_{-\infty}^{\infty} e\,\phi(e)\,a(e,\alpha)\,de + \frac{1}{2}s_{\theta\theta}(\alpha,\theta)\right) + O(n^{-(l+3)/2}), \tag{29}$$

where s_θ and $s_{\theta\theta}$ denote the first and second derivatives of $s(\alpha,\theta)$ with respect to θ. For given α, (29) is simply a function of θ.

There are two leading order terms in expression (29), and these are of order at most $n^{-3/2}$, as we would expect from (20). Neither of these terms depends on the level of s, the P value discrepancy function for the asymptotic test. Instead, they depend on s_θ, the slope of s, and on $s_{\theta\theta}$, which is a measure of the curvature of s. There is thus no reason to believe

that the performance of the bootstrap test will necessarily be worse for asymptotic tests that perform poorly than for asymptotic tests that perform well. At one extreme, the asymptotic test may be pivotal, in which case s will be flat, and (29) will then be zero no matter how poorly the asymptotic test performs. At the other extreme, there may well be cases in which s happens to be zero for a particular value of θ, so that the asymptotic test performs perfectly, and yet the bootstrap test will almost certainly not perform perfectly.

Let us now consider the two leading-order terms in (29). The first term is proportional to s_θ. The integral in it can readily be seen to be proportional to the bias of the estimator $\hat{\theta}$, conditional on α; recall (28). When this bias is nonzero, the bootstrap will, on average, be evaluating $Q(\alpha, \theta)$ at the wrong point. That will not matter if $S(\alpha, \theta)$ is flat, in which case $s_\theta = 0$, and the first term vanishes. However, it will matter if S is not flat. Suppose, for concreteness, that $s_\theta > 0$ and $E(\hat{\theta}) > \theta_0$, so that the first term in (29) is negative. In this case, the average of the $Q(\alpha, \hat{\theta})$ over the $\hat{\theta}$ will be less than $Q(\alpha, \theta_0)$; remember that Q is the inverse of S, and recall Figure 2. This means that the bootstrap test will not reject often enough, and its P value discrepancy must therefore be negative.

Even if $\hat{\theta}$ is unbiased, when S is nonlinear, so that its graph is curved and $s_{\theta\theta}$ is nonzero, then the curvature will lead to the average of the $Q(\alpha, \hat{\theta})$ being different from $Q(\alpha, \theta_0)$. For example, if $s_{\theta\theta}$ is negative, then $q_{\theta\theta}$ will be positive, and the average of the $Q(\alpha, \hat{\theta})$ will consequently be too large. This means that, in this case, the bootstrap test will reject too often, and its P value discrepancy will be positive.

Notice that if $\hat{\theta}$ is unbiased, at least to highest order, and if S is linear, then both the leading terms in (29) will vanish, and the bootstrap test will work perfectly, at least through $O(n^{-3/2})$. Even though the rejection region of the bootstrap test will be different from the true theoretical one whenever s_θ is not zero, as much probability mass will be gained on one side as is lost on the other in going from one region to the other; see Figure 1.

What have we learned in this section? We already knew, from the results of Section 3, and in particular (20), that the size distortion of the bootstrap test is, under plausible circumstances, a full order of magnitude smaller than the size distortion of the asymptotic test. What we have learned from (29) is that the size distortion of the bootstrap test depends in a particular way on the shape of the P value discrepancy function for the asymptotic test. If the bootstrap is based on unbiased parameter estimates, then only the curvature of this function matters. If it is based on biased parameter estimates, then the slope matters as well. In contrast, the level of the P value discrepancy function for the asymptotic test never matters.

Although (29) applies only to the one-parameter case, these results must evidently be true more generally.

6. Bootstrap Tests for Serial Correlation

In this section and the next one, we provide some Monte Carlo evidence on the performance of bootstrap tests. There are several interesting results, of which two stand out. The first is that, in many circumstances, bootstrap tests work extraordinarily well. The second is that the theory of Section 5 is very useful in understanding when bootstrap tests will and will not work well.

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

$$(30) \qquad y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Y}_t\boldsymbol{\delta} + u_t, \quad u_t = \sum_{l=1}^{r} \rho_l u_{t-l} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma^2),$$

where $\boldsymbol{X}_t$ is a $k \times 1$ vector of regressors that may be treated as fixed, $\boldsymbol{Y}_t$ 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 (30) under the null hypothesis so as to obtain estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$, and residuals $\hat{u}_t \equiv y_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}} - \boldsymbol{Y}_t\hat{\boldsymbol{\delta}}$, and then run the regression

$$(31) \qquad y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Y}_t\boldsymbol{\delta} + \sum_{l=1}^{r} \rho_l \hat{u}_{t-l} + \text{residual},$$

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

$$(32) \qquad \frac{n-k-m-r}{r} \times \frac{\|\boldsymbol{P}_{M_{[\boldsymbol{X}\ \boldsymbol{Y}]}\hat{\boldsymbol{V}}}\hat{\boldsymbol{u}}\|^2}{\|\boldsymbol{M}_{M_{[\boldsymbol{X}\ \boldsymbol{Y}]}\hat{\boldsymbol{V}}}\hat{\boldsymbol{u}}\|^2},$$

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

This approach, which is due principally to Durbin (1970) and Godfrey (1978), 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; see Kiviet (1986). However, the test statistic (32) is not exact in finite samples, and it is therefore natural to bootstrap it. The procedure is as follows:

1. Estimate (30) by OLS under the null hypothesis that $\rho_1 = \rho_2 = \ldots = \rho_r = 0$ so as to obtain $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\delta}}$, and $\hat{\boldsymbol{u}}$. Then construct $\hat{\boldsymbol{V}}$ from $\hat{\boldsymbol{u}}$ and compute the test statistic (32), which, following our earlier notation, we will call $\hat{\tau}$.

2. Draw B sets of bootstrap error terms, $\boldsymbol{u}^j$, and use them to generate B bootstrap samples $\boldsymbol{y}^j$. There are numerous ways in which the error terms can be drawn, four of which will be described below. The elements of $\boldsymbol{y}^j$ are generated recursively from the equation

$$y_t^j = \boldsymbol{X}_t\hat{\boldsymbol{\beta}} + \boldsymbol{Y}_t^j\hat{\boldsymbol{\delta}} + u_t^j, \tag{33}$$

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

3. For each bootstrap sample, compute τ_j, the test statistic (32) that uses $\boldsymbol{y}^j$ and $\boldsymbol{Y}^j$ instead of $\boldsymbol{y}$ and $\boldsymbol{Y}$. Then compute the estimated bootstrap P value as

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I(\tau_j > \hat{\tau}). \tag{34}$$

In the limit, as $B \to \infty$, this tends to the bootstrap P value $p^*(\hat{\tau})$; compare (3). In practice, as we discuss in Davidson and MacKinnon (1996), it may be desirable to use more complicated ways to estimate the bootstrap P value, but in this paper we use only the simplest possible estimator, which is (34).

We consider four different ways of generating the u_t^j. 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^j by resampling with replacement from the vector with typical element

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

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 recentered 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^j 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 $\boldsymbol{P}_{[\boldsymbol{X}\ \boldsymbol{Y}]}$. Then recenter 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 (32) can be written as

$$\frac{n-k-r}{r} \times \frac{\|\boldsymbol{P}_{\boldsymbol{M_X}\hat{\boldsymbol{V}}}\boldsymbol{M_X}\boldsymbol{u}\|^2}{\|\boldsymbol{M}_{\boldsymbol{M_X}\hat{\boldsymbol{V}}}\boldsymbol{M_X}\boldsymbol{u}\|^2}, \tag{35}$$

which depends only on the matrix $\boldsymbol{X}$ and the vector $\boldsymbol{u}$; recall that each column of $\hat{\boldsymbol{V}}$ is just $\boldsymbol{M_X}\boldsymbol{u}$ lagged some number of times. The only parameter that affects $\boldsymbol{u}$ is σ, and (35), like all F statistics, is invariant to its value. Thus (35) is pivotal and, by the result discussed in Section 2, the bootstrap P value is equal to the true P value for the test. Note that, in this case, step 2 can be simplified, since (33) is no longer needed; we can just generate the y_t^j as independent $N(0,1)$ random variables.

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 (a midrange PC at the time this is being written), 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 1000 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 (32) does implicitly depend

on all the parameters through the process that generates y_t recursively. However, the theoretical results of Section 3 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.

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, $j = 1, \ldots, 4$. 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 3 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 $\boldsymbol{X}$ 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.

We used Figure 3 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. As the theory of Section 5 makes clear, 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 (32) 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 $B = 1000$ 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. We chose not to use control variates based on asymptotic theory. The benefit from doing so would have been very small because most of the experiments involve very small sample sizes, and because we

are computing tail-area probabilities; see Davidson and MacKinnon (1992). It was necessary to use a fairly large value for B because, as we discuss in Davidson and MacKinnon (1996), bootstrap P values calculated as in (34) are usually biased. This bias is $O(1/B)$, and using $B = 1000$ ensures that it is negligible.

The key results of our experiments are presented in Figure 4. 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.

We now turn our attention to power. The results of Section 4 suggest that, on a size-corrected basis, the power of the bootstrap test should be very similar to the power of the t test. However, as we hinted in that section, 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 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 further 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 parametrized. 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 parametrization, 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 parametrization-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 parametrized.

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 parametrization-independent manner and of introducing no unnecessary dependence on the sample size. Moreover, Horowitz (1995) 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 (30), 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 $L(1-\delta_1 L)^{-1}\boldsymbol{X}\boldsymbol{\beta}_1$ on $\boldsymbol{X}$ and define $\boldsymbol{b}_2$ as the vector of parameter estimates and S as $1/n$ times the sum of squares of the residuals from that regression. 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

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

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

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

between the two measures of power for the bootstrap tests. The bootstrap test results here are always for 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 of Section 4 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 3. 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 of Section 4 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.

In addition to confirming the theoretical results of Sections 3 through 5, 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.

7. 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; see, among others, Godfrey and Pesaran (1983). It therefore seems natural to bootstrap the J test.

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

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

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

$$\boldsymbol{y} = \boldsymbol{Xb} + \alpha \boldsymbol{P_Z y} + \text{residuals}, \tag{36}$$

where $\boldsymbol{P_Z} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top$. Thus $\boldsymbol{P_Z y}$ 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 (36) 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_j, and we then compute the estimated bootstrap P value via equation (34). 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 in the last section.

If $\acute{s}$ denotes the estimated standard error from regression (36), the J test statistic can be written as

$$\frac{\boldsymbol{y}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{y}}{\acute{s}(\boldsymbol{y}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{P}_Z \boldsymbol{y})^{1/2}}, \tag{37}$$

where $\boldsymbol{M}_X \equiv \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. It is straightforward to show that, under H_1, the statistic (37) depends on both $\boldsymbol{\beta}$ and σ_1, but only through the ratio $\boldsymbol{\beta}/\sigma_1$. Thus, if we choose a fixed vector $\boldsymbol{\beta}^*$ and let $\boldsymbol{\beta} = \delta\boldsymbol{\beta}^*$, 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. 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$. Each of the columns of $\boldsymbol{X}$, except for the constant term, was made up of i.i.d. normal random variables, was independent of the other columns, and was normalized to have length n. Each column of $\boldsymbol{Z}$ was correlated with one of the columns of $\boldsymbol{X}$, with squared correlation 0.5 in most of our experiments. All elements of $\boldsymbol{\beta}^*$ were equal.

Figure 5 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 $\boldsymbol{Z}$ that are not in $\boldsymbol{X}$; compare Figure 2, which is for the case with $k_1 = 2$ and $k_2 = 4$, for $n = 25$. For the smaller sample sizes, the performance of the 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 theory of Section 3 to apply either.

On the basis of Figure 5, 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 6 may therefore come as a surprise. This figure, which is similar to Figure 4, 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 $B = 1000$.

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 5 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.

These results provide strong support for the theory of Sections 3 and 5. 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. Of course, we cannot claim on the basis of these results that bootstrap J tests will *always* work well. There undoubtedly exist situations in which PVFs are even steeper or more sharply curved than the ones in Figure 5, and for which bootstrap tests consequently work less well. It is certainly necessary to stay away from situations in which the underlying asymptotic theory does not hold.

8. Summary and Conclusion

In this paper, we have advocated the use of the bootstrap in many hypothesis testing situations where exact tests are not available. In particular, we have advocated the use of bootstrap P values, because P values are more informative than the reject/do-not-reject results of tests with some pre-chosen size, and the actual calculation of bootstrap P values is, if anything, easier than the calculation of bootstrap critical values. In addition, the theory of bootstrap P values, as presented in this paper, is no more difficult than the theory of bootstrap critical values.

The bootstrap provides higher-order refinements, relative to asymptotic theory, whenever the quantity bootstrapped is, asymptotically at least, pivotal. This is the case for all commonly used test statistics in econometrics. As we discussed in Section 3, 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}.

The results of Section 3 can be applied to any bootstrap test of level α whenever we know the order of magnitude of the bias of the α quantile of the bootstrap distribution of the statistic considered as an estimator of the α quantile of the true distribution of the statistic. In Section 5, we obtained more detailed results, which, strictly speaking, apply only to the case of the parametric bootstrap applied to a fully specified model. However, even "nonparametric" bootstrap distributions 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 in Sections 6 and 7. 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.

The P value discrepancy function is central to the results of Section 5. For given nominal size, this function measures, as a function of the actual DGP, the extent to which the actual size differs from the nominal size. 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 $\boldsymbol{\theta}$. 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 $\boldsymbol{\theta}$. 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 level of the P value discrepancy function is the size distortion of the asymptotic test. However, this level 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.

Section 4 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 shown that, once both tests are corrected for size, a bootstrap test can have different power from the corresponding asymptotic test only if the bootstrap distribution has different properties under the null and under the alternative. For the parametric bootstrap, the bootstrap distribution satisfies the null hypothesis in all cases. Thus, power differences can arise only from differences in the behavior of the parameter estimates when they are estimating the parameters of a DGP that actually satisfies the null and when they are estimating the pseudo-true parameters of a DGP that does not.

It is important to stress the fact that, although the size distortions of bootstrap tests that we have studied in this paper 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 misbehavior, so as to provide confirmation of our 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.

Appendix

In this appendix, we derive expression (29) for the P value discrepancy of the bootstrap test in the one-parameter case. First, the function Q is redefined to take as second argument a parameter θ rather than a DGP μ. Thus γ, σ_γ, and k are defined by

$$n^{-k/2}\sigma_\gamma\gamma = Q(\alpha, \theta_0 + n^{-1/2}e) - Q(\alpha, \theta_0) \tag{A.01}$$

and by the requirement that the variance of γ should asymptotically be unity; see (14).

The relation between γ and e can be obtained to desired order from (A.01). We use the analogue of (10) in order to define the integer l and the function $q(\alpha, \theta)$, and then obtain by Taylor expansion:

$$\begin{aligned} n^{-k/2}\sigma_\gamma \gamma = n^{-(l+1)/2}\, e\, q_\theta(\alpha, \theta_0) + \tfrac{1}{2} n^{-(l+2)/2} e^2 q_{\theta\theta}(\alpha, \theta_0) \\ + O(n^{-(l+3)/2}), \end{aligned} \tag{A.02}$$

where q_θ and $q_{\theta\theta}$ denote the first and second derivatives of $q(\alpha, \theta)$ with respect to θ. Since the variance of e is 1 by assumption, we see directly from (A.02) that $k = l + 1$, and that

$$\sigma_\gamma = |q_\theta(\alpha, \theta_0)|. \tag{A.03}$$

We can also see from (A.02) that γ and e are equal to leading asymptotic order. Thus the function $h(\gamma)$ of (19) is just $\phi(\gamma)$. Note that it is enough to define σ_γ as in (A.03) in such a way that γ has variance unity to leading order asymptotically, since any discrepancy at lower order can be caught in the function f in (19).

Let us assume for simplicity that $q_\theta(\alpha, \theta_0)$ is positive. Then, after removing unnecessary powers of n and using (A.03), (A.02) becomes

$$\gamma = e + \tfrac{1}{2} n^{-1/2} e^2 \frac{q_{\theta\theta}(\alpha, \theta_0)}{q_\theta(\alpha, \theta_0)} + O(n^{-1}). \tag{A.04}$$

This relationship may be inverted so as to express e in terms of γ:

$$e = \gamma - \tfrac{1}{2} n^{-1/2} \gamma^2 \frac{q_{\theta\theta}(\alpha, \theta_0)}{q_\theta(\alpha, \theta_0)} + O(n^{-1}). \tag{A.05}$$

In order to implement (20), we also need an expression for $f(\gamma, \alpha)$ valid at least to leading order. For this, we must use the information in (A.05) over and above the simple asymptotic equality of e and γ. We wish to find the density of γ conditional on $\pi_\tau = \alpha$. The density of e conditional on π_τ is just the product of the last two factors in (28). Thus, since e and γ are related, without reference to π_τ, by (A.04) and (A.05), the density of γ conditional on $\pi_\tau = \alpha$ is just

$$\phi(e)\big(1 + n^{-1/2} a(e, \alpha)\big)\, \frac{de}{d\gamma}. \tag{A.06}$$

where e is related to γ by (A.05), from which we compute

$$\frac{de}{d\gamma} = 1 - n^{-1/2}\gamma\,\frac{q_{\theta\theta}(\alpha,\theta_0)}{q_\theta(\alpha,\theta_0)} + O(n^{-1}).$$

Thus (A.06) becomes

$$\begin{aligned}\phi\big(\gamma - \tfrac{1}{2}n^{-1/2}\gamma^2\frac{q_{\theta\theta}}{q_\theta} + O(n^{-1})\big)\big(1 + n^{-1/2}a(\gamma,\alpha) + O(n^{-1})\big)\\ \times\big(1 - n^{-1/2}\gamma\,\frac{q_{\theta\theta}}{q_\theta} + O(n^{-1})\big),\end{aligned} \tag{A.07}$$

where $q_{\theta\theta}$ and q_θ without explicit arguments are evaluated at (α,θ_0). In order to simplify this expression, note that

$$\phi\Big(\gamma - \tfrac{1}{2}n^{-1/2}\gamma^2\frac{q_{\theta\theta}}{q_\theta} + O(n^{-1})\Big) = \phi(\gamma)\Big(1 + \tfrac{1}{2}n^{-1/2}\gamma^3\frac{q_{\theta\theta}}{q_\theta} + O(n^{-1})\Big).$$

Thus, to leading order, (A.07) simplifies to

$$\phi(\gamma)\left(1 + n^{-1/2}\frac{q_{\theta\theta}}{q_\theta}\big(\tfrac{1}{2}\gamma^3 - \gamma\big) + n^{-1/2}a(\gamma,\alpha)\right). \tag{A.08}$$

If we had not assumed that $j = 1$, the factor in front of the third term inside the large parentheses would have been $n^{-j/2}$, and this term would not have been of leading order for $j > 1$. Comparing (A.08) with (19) shows that, in the latter of these expressions,

$$f(\gamma,\alpha) = \frac{q_{\theta\theta}}{q_\theta}\big(\tfrac{1}{2}\gamma^3 - \gamma\big) + a(\gamma,\alpha).$$

We may finally return to (20), and substitute in all the results we have obtained for this special case. The integral will be written with dummy variable e rather than γ, since the two random variables e and γ are asymptotically equal, and since it is clearer intuitively to reason in terms of e, which is $n^{1/2}$ times the estimation error in $\hat\theta$, rather than γ. The size distortion of the bootstrap test is then

$$n^{-(l+2)/2}q_\theta\int_{-\infty}^{\infty} de\; e\,\phi(e)\left(\frac{q_{\theta\theta}}{q_\theta}\big(\tfrac{1}{2}e^3 - e\big) + a(e,\alpha)\right) + O(n^{-(l+3)/2}).$$

Since the fourth moment of the standard normal distribution equals 3, the above expression is

$$n^{-(l+2)/2}\left(\tfrac{1}{2}q_{\theta\theta} + q_\theta\int_{-\infty}^{\infty} de\; e\,\phi(e)\,a(e,\alpha)\right) + O(n^{-(l+3)/2}). \tag{A.09}$$

Alternatively, (A.09) may be expressed in terms of the derivatives of the P value discrepancy function s, evaluated at (α,θ_0). Except for a sign change, the result is essentially the same:

$$-n^{-(l+2)/2}\left(\tfrac{1}{2}s_{\theta\theta} + s_\theta\int_{-\infty}^{\infty} de\; e\,\phi(e)\,a(e,\alpha)\right) + O(n^{-(l+3)/2}). \tag{A.10}$$

Expression (29) of the text is simply (A.10) rewritten slightly.

References

Attfield, C. L. F. (1995). "A Bartlett adjustment to the likelihood ratio test for a system of equations," *Journal of Econometrics*, 66, 207–223.

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, 1305–1329.

Davidson, R. and J. G. MacKinnon (1992). "Regression-based methods for using control variates in Monte Carlo experiments," *Journal of Econometrics*, 54, 1992, 203–222.

Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, Oxford University Press.

Davidson, R. and J. G. MacKinnon (1996). "Computing bootstrap tests," manuscript.

Durbin, J. (1970). "Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables," *Econometrica*, 38, 410–421.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*, New York, Chapman and Hall.

Godfrey, L. G. (1978). "Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables," *Econometrica*, 46, 1303–1310.

Godfrey, L. G., and M. H. Pesaran (1983). "Tests of non-nested regression models: small sample adjustments and Monte Carlo evidence," *Journal of Econometrics*, 21, 133–154.

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

Hall, P. and D. M. Titterington (1989). "The effect of simulation order on level accuracy and power of Monte-Carlo tests," *Journal of the Royal Statistical Society*, Series B, 51, 459–467.

Horowitz, J. L. (1994). "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, 61, 395–411.

Horowitz, J. L. (1995). "Bootstrap methods in econometrics: Theory and numerical performance," paper presented at the 7th World Congress of the Econometric Society, Tokyo.

Kiviet, J. F. (1986). "On the rigour of some misspecification tests for modelling dynamic relationships," *Review of Economic Studies*, 53, 241–261.

Rothenberg, T. J. (1984). "Hypothesis testing in linear models when the error covariance matrix is nonscalar," *Econometrica*, 52, 827–842.

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.

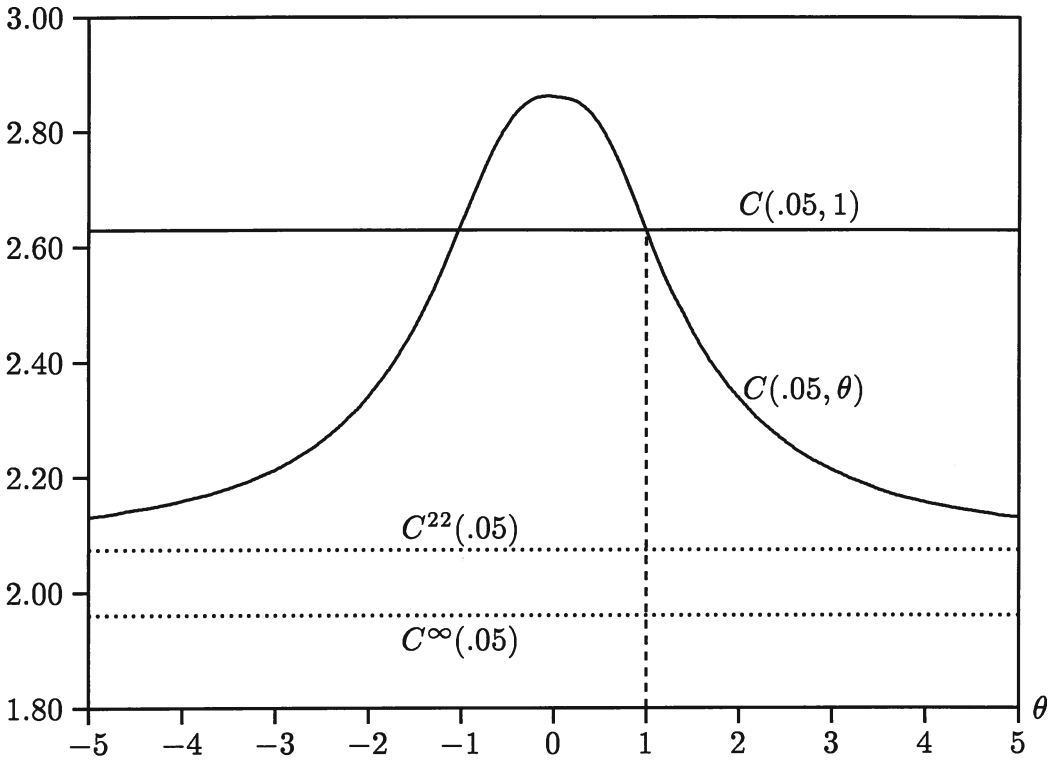

Figure 1. A Critical Value Function

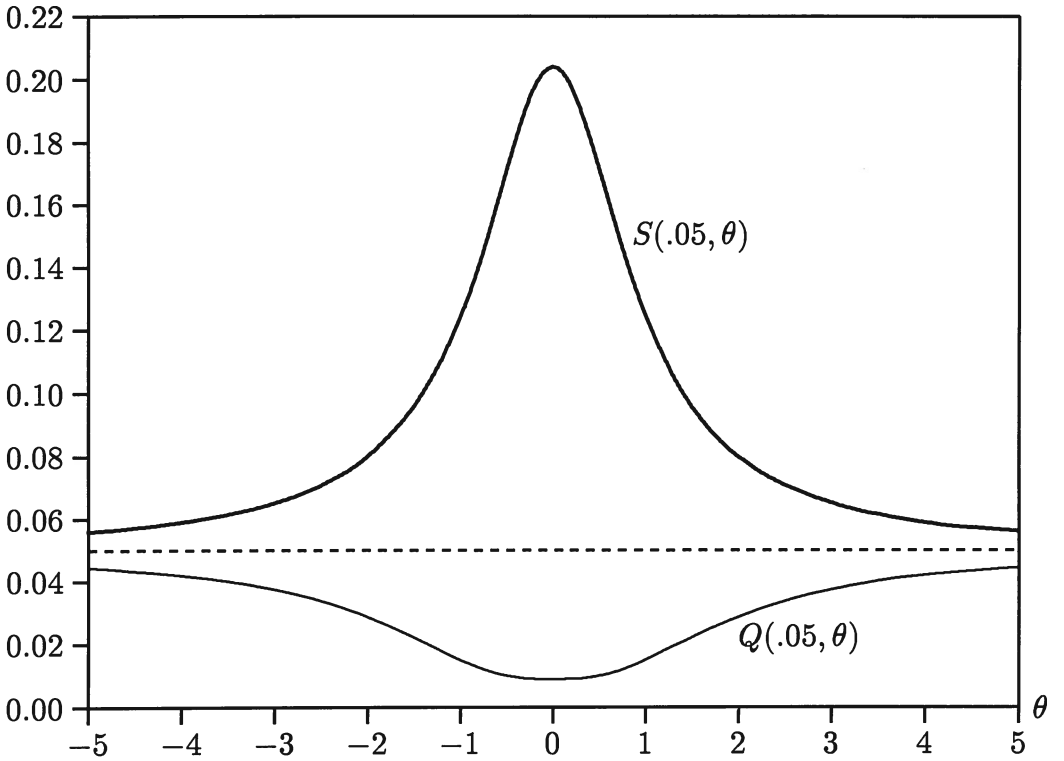

Figure 2. A P Value Function and its Inverse

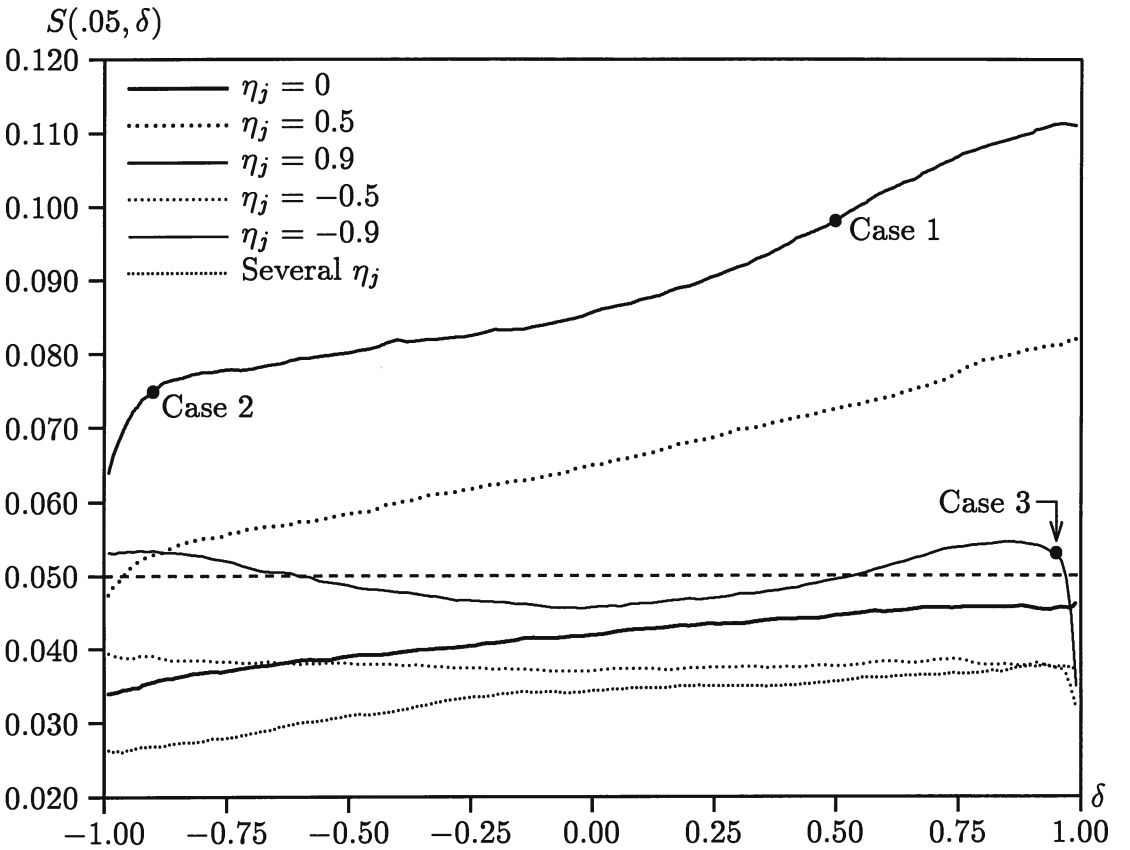

Figure 3. P Value Functions for AR(1) Tests at .05 Level, $n = 25$

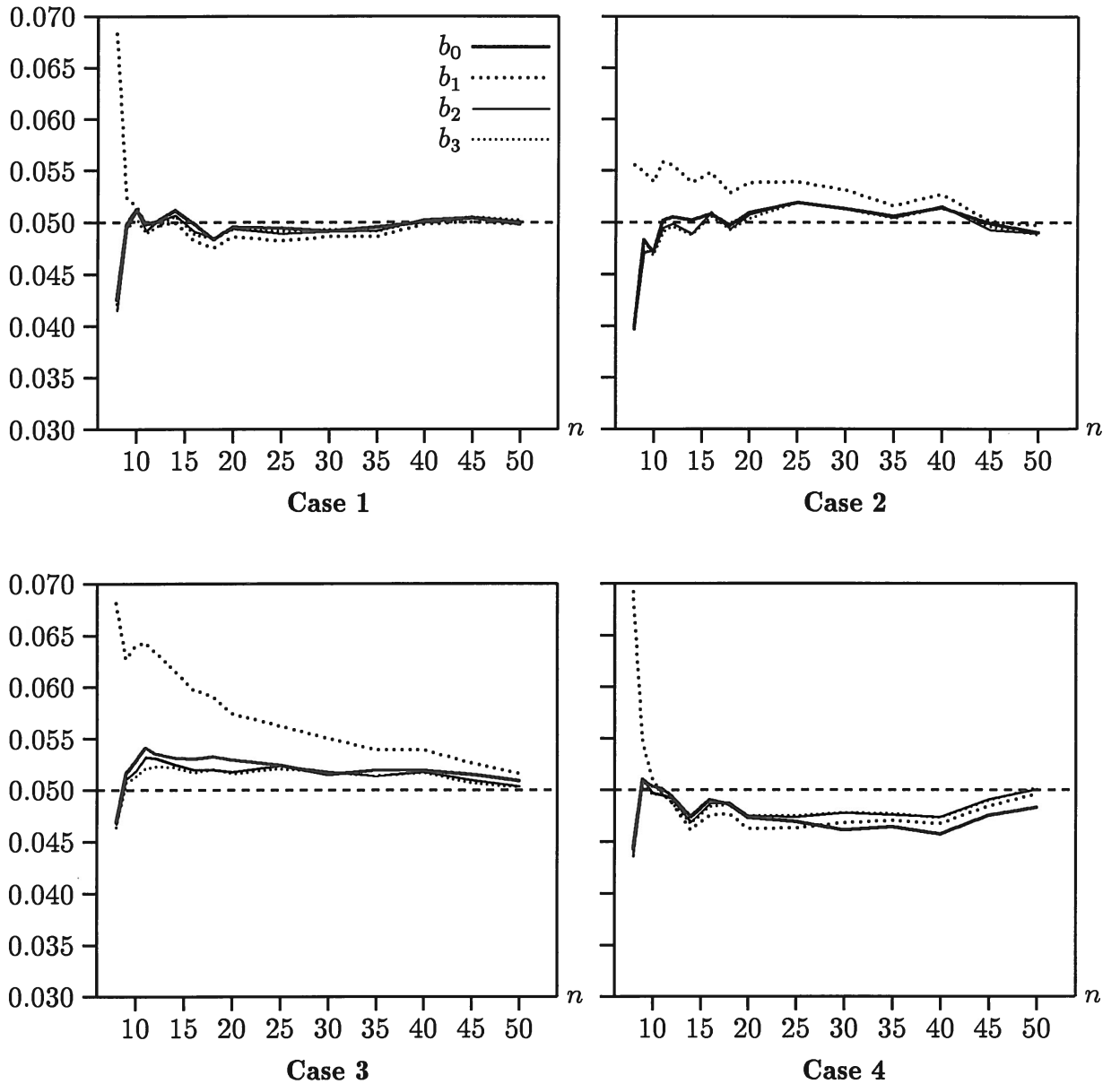

Figure 4. Estimated P Values for Bootstrap AR(1) Tests at .05 Level

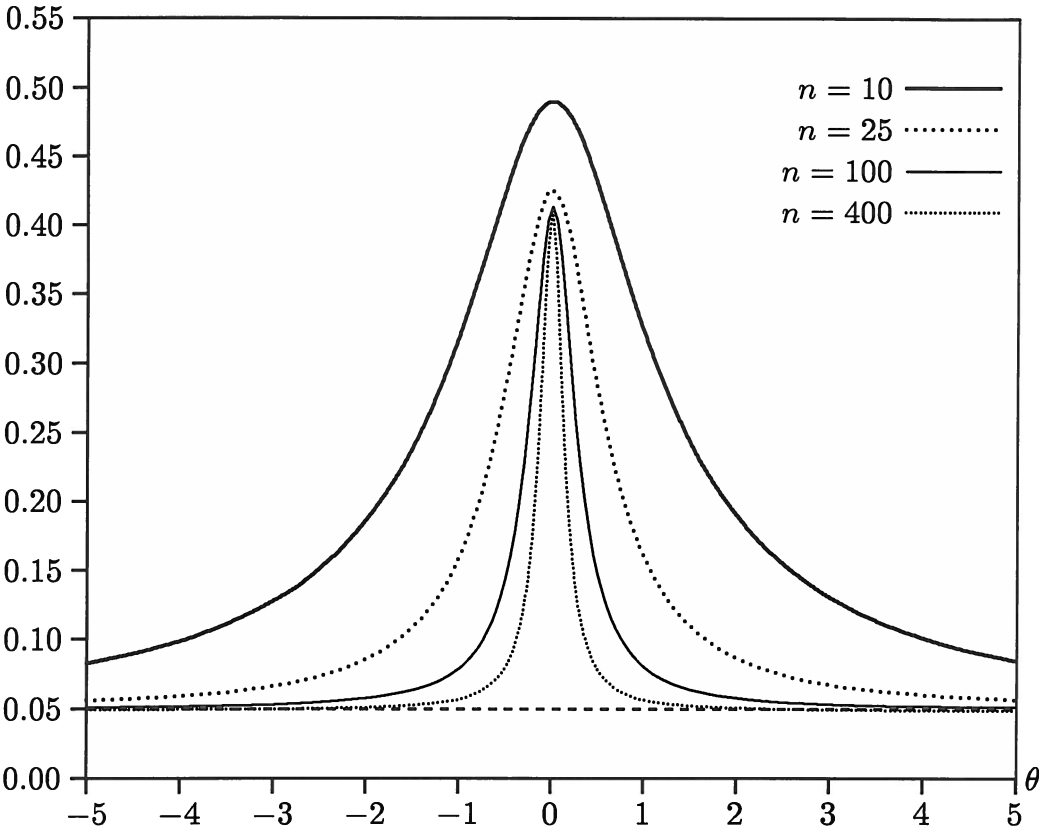

Figure 5. P Value Functions for J Tests

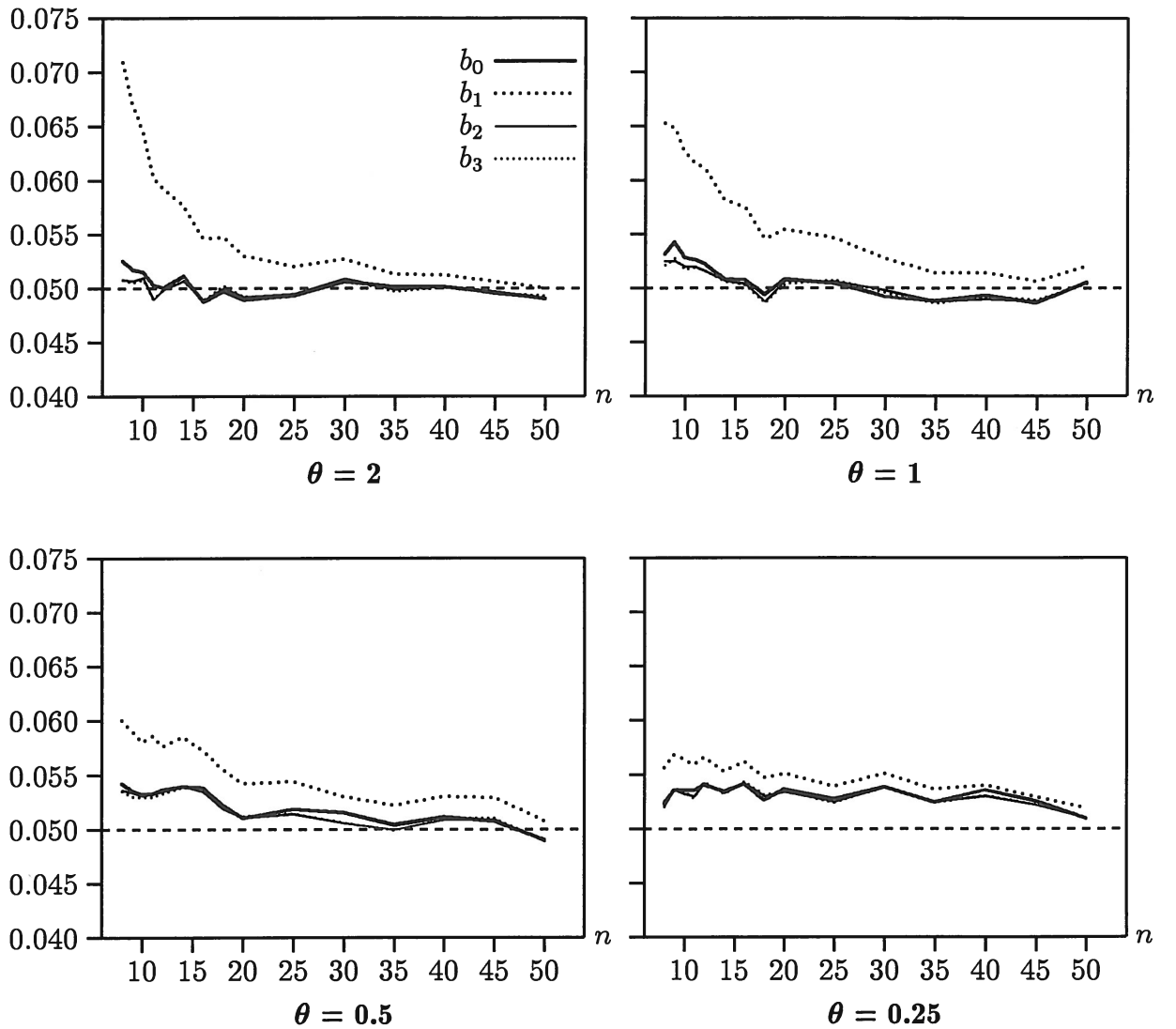

Figure 6. Estimated P Values for Bootstrap J Tests at .05 Level

The Size and Power of Bootstrap Tests

by

Russell Davidson

and

James G. MacKinnon