Queen’s Economics Department Working Paper No. 936

The Size Distortion of Bootstrap Tests

Russell Davidson
GREQAM, Queen’s University

James G. MacKinnon
Queen’s University

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

08-1996

The Size Distortion 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

Russell Davidson: **russell@ehess.cnrs-mrs.fr**
James MacKinnon: **jgm@qed.econ.queensu.ca**

Abstract

Bootstrap tests are tests for which the significance level is calculated by some sort of bootstrap procedure, which may be parametric or nonparametric. We provide a theoretical framework in which to study the size distortions of bootstrap P values. We show that, in many circumstances, the size distortion of a bootstrap test will be one whole order of magnitude smaller than that of the corresponding asymptotic test. 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. Monte Carlo results are presented for the case of nonnested hypothesis tests. These results confirm and illustrate the utility of our theoretical results, and they also suggest that bootstrap tests may often work extremely well in practice.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. An earlier version was presented at Universidad Carlos III de Madrid, Universidad Complutense de Madrid, Cambridge University, INSEE-CREST (Paris), CORE (Louvain-la-Neuve), the Tinbergen Institute (Amsterdam), the University of Geneva, the European University Institute (Florence), the ESRC Econometrics Conference (Bristol), and the 1996 Berkeley Symposium on the Bootstrap. We are grateful to many seminar participants and to two anonymous referees for comments on the earlier version. We are especially grateful to Joel Horowitz, not only for comments, but also for his probing questions that led us to clarify the paper. The paper was written while the second author was visiting GREQAM.

August, 1996

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 in econometrics, 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 or regions rather than test statistics; see, among many others, Efron and Tibshirani (1993) and Hall (1992). Many authors consider this point of little importance. In particular, Hall, in the only section of the book just cited in which he discusses hypothesis testing at any length, points out that there “is a well-known and established duality between confidence intervals and hypothesis testing.” On the other hand, in a survey paper, Hinkley (1988) did briefly mention matters related to significance tests and the bootstrap. More importantly, Beran (1986) drew a distinction between the confidence region approach that has continued to be dominant in the statistical literature, and what he called a test statistic approach. The distinction is valuable in many econometric contexts, because conventional confidence intervals, based on unrestricted estimation of the parameters with which the null hypothesis under test is concerned, are dual only to Wald tests. It is well known that Wald tests have the defect that they are not invariant under nonlinear reparametrizations of the restrictions under test; see Chapter 13 of Davidson and MacKinnon (1993) for a discussion and references. In fact, not unexpectedly in view of the above-mentioned duality, bootstrap percentile-t confidence intervals, although their properties are usually better than those of other bootstrap-based confidence intervals, share this defect with Wald tests; see Hall (1992).

It is perfectly possible to bootstrap any sort of test, not just Wald tests. The basic idea of bootstrapping any test statistic is to draw a large number of “bootstrap samples” from a distribution which obeys the null hypothesis and is constructed in such a way that the bootstrap samples, as far as possible, resemble the real sample. One then compares the observed test statistic with the ones calculated from the bootstrap samples. An important recent paper which advocates this approach is Horowitz (1994).

In this paper, we point out that bootstrapping many of the commonly used test statistics in econometrics leads to drastic improvements in the reliability of inference compared with conventional asymptotic tests. In some cases, bootstrap inference is exact, while in others it is very nearly so. In many cases, it is very easy to implement what we shall call "bootstrap tests," that is, tests for which the significance level is calculated by some sort of bootstrap procedure. Despite this, there is little discussion of bootstrap testing methodology in the econometrics literature. A noteworthy recent exception is Hall and Horowitz (1996), in which bootstrap methods are described for dynamic GMM models.

The bootstrap's main virtue in hypothesis testing is that, when properly used, it leads to asymptotic refinements in significance levels. This means that the error in the P value given by a bootstrap test is of lower order in the sample size n than the error given by a corresponding asymptotic test. For an asymptotic test, the error in the P value is in general of order $n^{-1/2}$. Depending on the circumstances, use of the bootstrap can reduce this order to n^{-1}, $n^{-3/2}$, or even smaller values; see Hall (1992) for a very full discussion, based on Edgeworth expansions, of the extent to which asymptotic refinements are available in different contexts. In this paper, we make no explicit use of Edgeworth expansions, and instead consider all test statistics in approximate P value form. This device allows us to deal conveniently with all the possible asymptotic distributions that test statistics may have, and thereby to provide a unifying theoretical framework in which the determinants of the order of refinements can be studied. This leads to the discovery of a new refinement which will often be available in econometric applications.

Even when it is known theoretically that the bootstrap provides refinements of a given order, the size distortions of bootstrap tests — the errors in the P values it provides — may vary considerably in different circumstances. In this paper, we provide an explanation of this phenomenon in terms of what we call "critical value functions" or "P value functions." We show how the slope and curvature of these functions affects bootstrap size distortions, and we provide examples in which graphs of the functions depict clearly those regions in the parameter space where the bootstrap will behave more or less well. Inspection of such graphs can yield valuable intuition concerning bootstrap tests.

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 and 4. In Section 5, we present Monte Carlo results, on the J test for nonnested hypotheses, which confirm and illustrate the utility of our theoretical results.

2. Basic Concepts and Notation

Suppose that we calculate a test statistic $\hat{\tau}$ from a sample of size n. The details of how $\hat{\tau}$ is calculated need not concern us, but it is essential for the asymptotic refinements of the bootstrap that $\hat{\tau}$ be asymptotically pivotal. In other words, the asymptotic distribution of $\hat{\tau}$ under the null hypothesis must not depend on any unknown parameters, or, more generally, on just which data-generating process (DGP) belonging to the null hypothesis actually generated the data. 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. The importance of pivotalness in bootstrap methodology has been insisted on by many authors. Hall (1992) gives extensive bibliographical notes on this point; see especially the ends of his Chapters 1 and 3.

It is possible to use bootstrapping either to calculate a critical value for $\hat{\tau}$ or to calculate the significance level, or P value, associated with it. 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 this paper.

Suppose that the data from which $\hat{\tau}$ was calculated were generated by a DGP which actually satisfies the null hypothesis. We denote this DGP by μ_0. Then, for a one-tailed test, the P value we would ideally like to compute is

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

where τ denotes the random variable of which $\hat{\tau}$ is a realization. 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.

Since μ_0 is unknown, we cannot compute (1). However, we can use the bootstrap to estimate it. We will define the *bootstrap P value* as

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

The only difference between (2) and (1) is that the former uses a DGP $\hat{\mu}$, which we will call a *bootstrap DGP*, instead of the actual DGP μ_0 to compute the probability. The bootstrap DGP may be obtained from either a parametric or a nonparametric bootstrap. In the former case, it is a DGP 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, the bootstrap DGP will be based on 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 Section 5, parameter estimates are often required to implement nonparametric bootstrap procedures. In either case, the bootstrap DGP will depend on the sample used to obtain $\hat{\tau}$.

In practice, the bootstrap P value (2) will be approximated by Monte Carlo sampling from the bootstrap DGP $\hat{\mu}$. Since it is often feasible to make the approximation so good that the approximation error can safely be ignored, we shall in this paper be concerned solely with theoretical bootstrap P values based on (2). Our objective is to understand the relationship between $p^*(\hat{\tau})$ and $p(\hat{\tau})$.

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 $\hat{\tau}$ is exactly pivotal, provided of course that the parametric model is correct. In this case, the only difference between the bootstrap DGP $\hat{\mu}$ and μ_0 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 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 many tests for serial correlation, ARCH errors and other forms of heteroskedasticity, skewness, and kurtosis, including the information matrix test; see, for instance, 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 $\hat{\tau}$ 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. Thus, in order to understand the properties of bootstrap P values, we need to specify how the distribution of $\hat{\tau}$ depends on μ. Suppose, for simplicity, that we are concerned with a one-tailed test which will reject the null hypothesis when $\hat{\tau}$ is in the upper tail. We therefore define the *critical value function*, or CVF, $Q(\alpha, \mu)$, by the equation

$$\text{(3)} \qquad \Pr_\mu\big(\hat{\tau} \geq Q(\alpha, \mu)\big) = \alpha.$$

$Q(\alpha, \mu)$ is thus the true level-α critical value for a one-tailed test based on the statistic $\hat{\tau}$ if the DGP is μ. In other words, it is the $1 - \alpha$ quantile of the distribution of $\hat{\tau}$ under μ. The rejection region for the bootstrap test of nominal size α is defined by

$$\text{(4)} \qquad \hat{\tau} \geq Q(\alpha, \hat{\mu}),$$

and the true size of the bootstrap test is simply the probability, under the DGP μ_0, of the event (4). This probability clearly depends only on the joint distribution under μ_0 of $\hat{\tau}$ and the scalar quantity $Q(\alpha, \hat{\mu})$. We shall make much use of this fact in the next two sections. Of course, we can easily redefine $Q(\alpha, \mu)$ to handle two-tailed tests or one-tailed tests with a rejection region in the lower tail, or, as we do shortly, to handle tests based on approximate P values.

In the parametric case, the DGP μ will belong to a parametric model and will be completely characterized by a parameter vector $\boldsymbol{\theta}$. Thus 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 two-tailed J test; see Section 5) with DGP characterized by a single parameter θ for $\alpha = .05$. From (4), the size of the bootstrap test is the probability, under θ_0, that $\hat{\tau} \geq Q(\alpha, \hat{\theta})$. This inequality defines a rejection region in the space of $\hat{\theta}$ and $\hat{\tau}$, namely, the region above the graph of the CVF. If the realized $(\hat{\theta}, \hat{\tau})$ falls into this region, the bootstrap test rejects.

The rectangle above the horizontal line marked $Q(.05, 1)$ 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 $Q^{\infty}(.05) = 1.96$ and using the critical value $Q^{22}(.05) = 2.074$ 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 overreject somewhat when $\hat{\theta} > 1$ and when $\hat{\theta} < -1$. For those values of θ, the CVF is below $Q(.05, 1)$, 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 4. In fact, as we shall see in Section 5, the bootstrap J test does work very well.

Suppose that the asymptotic distribution of $\hat{\tau}$ has c.d.f. F. The one-tailed asymptotic test based on $\hat{\tau}$ rejects at nominal level α if $1 - F(\hat{\tau}) \leq \alpha$. In finite samples, of course, the event $(1 - F(\hat{\tau}) \leq \alpha)$ will rarely have probability precisely α. Consequently, we introduce the *P value function*, or PVF, defined as follows:

$$S(\alpha, \mu) \equiv \Pr_{\mu}\big(1 - F(\hat{\tau}) \leq \alpha\big). \tag{5}$$

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

$$S(\alpha, \mu) = \alpha + n^{-l/2} s(\alpha, \mu), \tag{6}$$

where the integer $l \geq 1$ is defined so that $s(\alpha, \mu)$ will be $O(1)$. In the most general case, we expect that $l = 1$, but there are many exceptions. The most important of these is probably the case of a two-sided test based on an asymptotically $N(0, 1)$ statistic, or else an asymptotically chi-squared statistic. In this case, as Hall (1992) shows in Section 3.5.5, $l = 2$. Where no ambiguity is possible, we will often refer to the function s itself as the P value discrepancy function. Note that $s(\alpha, \mu)$ will be independent of the DGP μ if and only if $\hat{\tau}$ is pivotal.

The relation between the PVF S and the CVF Q can be expressed as

$$S\bigl(1 - F(Q(\alpha, \mu)), \mu\bigr) = \alpha, \tag{7}$$

a result which follows from evaluating (5) at $\alpha = 1 - F(Q(\alpha, \mu))$, rearranging, and using the definition (3) of $Q(\alpha, \mu)$.

The relation (7) becomes more transparent if $\hat{\tau}$, rather than being a general statistic, is a statistic in approximate P value form. In the present instance, the original $\hat{\tau}$ would be replaced by the asymptotic P value $1 - F(\hat{\tau})$, of which the asymptotic distribution is uniform on $[0, 1]$. If τ is thus redefined, (3) should be modified to take account of the fact that one usually rejects when a P value is less, rather than greater, than a given value. Thus, for statistics that are asymptotic P values,

$$\Pr_\mu\bigl(\hat{\tau} \leq Q(\alpha, \mu)\bigr) = \alpha. \tag{8}$$

It is clear that $Q(\alpha, \mu)$ is now the α quantile of the distribution of $\hat{\tau}$ under μ. Similarly, the definition (5) of S simplifies in this case to

$$S(\alpha, \mu) \equiv \Pr_\mu(\hat{\tau} \leq \alpha).$$

so that, for fixed μ, $S(\alpha, \mu)$ is just the c.d.f. of $\hat{\tau}$ evaluated at α. In addition, (7) becomes simply

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

and so the CVF $Q(\alpha, \mu)$ can in this case be thought of as the inverse of the P value function $S(\alpha, \mu)$. From (9) we also obtain

$$Q\bigl(S(\alpha, \mu), \mu\bigr) = \alpha,$$

since both S and Q are increasing in their first arguments. Analogously to (6), 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 (6). 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, after the original test statistic has been converted to an asymptotic P value. Both 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 its inverse the

PVF, 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 two sections a little bit difficult.

Before we move on to the next section, we present the standard analysis of why bootstrap tests based on pivotal test statistics perform better than asymptotic tests. To the best of our knowledge, this analysis first appeared in Beran (1988) for the case of the parametric bootstrap, which we discuss first. Suppose that the DGP μ is fully characterized by a parameter vector $\boldsymbol{\theta}$, so that the P value function can be written as $S(\alpha, \boldsymbol{\theta})$. From (6), the difference between the probability that $\hat{\tau} \leq \alpha$ according to the bootstrap DGP and the true probability is

$$S(\alpha, \hat{\boldsymbol{\theta}}) - S(\alpha, \boldsymbol{\theta}_0) = n^{-l/2}\big(s(\alpha, \hat{\boldsymbol{\theta}}) - s(\alpha, \boldsymbol{\theta}_0)\big). \tag{11}$$

By Taylor expanding s around $\boldsymbol{\theta}_0$, we obtain

$$S(\alpha, \hat{\boldsymbol{\theta}}) - S(\alpha, \boldsymbol{\theta}_0) \overset{a}{=} n^{-l/2} \boldsymbol{s}_{\boldsymbol{\theta}}^{\top}(\alpha, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{12}$$

where $\boldsymbol{s}_{\boldsymbol{\theta}}(\alpha, \boldsymbol{\theta}_0)$ 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 (12) is of order $n^{-(l+1)/2}$. Thus the bootstrap approximation to the distribution of $\hat{\tau}$ is in error only at order $n^{-(l+1)/2}$, better than the error of the asymptotic uniform distribution by a factor of $n^{-1/2}$.

In the case of the nonparametric bootstrap, Hall has shown that, in most cases, pivotal test statistics with an asymptotically standard normal distribution allow Edgeworth expansions of their distribution functions, equally well for the asymptotic statistic and for its bootstrap counterpart. Since the coefficients in such Edgeworth expansions are functions of the moments of the data generated by a DGP μ, and since sample moments, generated by $\hat{\mu}$, are root-n consistent estimators of the population moments, generated by μ_0, it follows that $s(\alpha, \hat{\mu}) - s(\alpha, \mu_0)$ is once more at most of order $n^{-(l+1)/2}$. The most convenient reference for this result is Hall (1992).

The above arguments are perfectly correct as far as they go, but they do not exploit the fact that $\hat{\tau}$ and $\hat{\mu}$ have a joint distribution. As we shall see in the next section, when this is taken into account, it turns out that the bootstrap test may perform even better, in many frequently encountered cases, than the simple analysis above suggests.

3. The Size of Bootstrap Tests

In this section, we obtain approximate expressions for the P value discrepancy function for bootstrap tests. These expressions usually have clear intuitive interpretations, based on the joint distribution of the test statistic $\hat{\tau}$ and the bootstrap DGP $\hat{\mu}$. We do not explicitly make use of Edgeworth expansions in our development, and so our results apply in a context at once more general and more specific than that treated in, for instance, Hall and Titterington (1989) and Hall (1992), where the results are based on Edgeworth expansions of statistics that are asymptotically standard normal or chi-squared. The generality of our results is that they apply to all sorts of hypothesis testing situations, and their specificity is that they apply only to hypothesis *tests*, rather than to confidence intervals. The duality between testing and confidence intervals presumably means that our results have counterparts in the confidence interval approach, but we will not pursue that issue here.

When it applies, the Edgeworth expansion approach is usually complementary to ours. Edgeworth expansions are usually effective for determining the order of asymptotic refinements in a given situation, and, if recast in our notation, they can provide analytical expressions for many of the quantities we study below.

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, as we did in the last section. The bootstrap critical value for $\hat{\tau}$, $Q(\alpha,\hat{\mu})$, is then 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 $\hat{\tau}$ 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:

$$Q(\alpha,\hat{\mu}) = Q(\alpha,\mu_0) + n^{-k/2}\sigma_\gamma\gamma, \tag{13}$$

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 (13) true.

In the case of the parametric bootstrap based on root-n consistent estimates, we find, by the standard argument given at the end of the last section, that $k = l + 1$. This is also true for the nonparametric bootstrap in those cases for which Hall's Edgeworth expansion theory applies. Recall from (12) that the difference between $S(\alpha,\hat{\mu})$ and $S(\alpha,\mu_0)$ in these cases 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 $k \geq 2$ in most cases of interest. Clearly, what is needed is that we should be able to write (10) not only for μ strictly satisfying the null hypothesis, but also for the discrete-valued DGPs $\hat{\mu}$ that are used as nonparametric bootstrap distributions. Then, provided that $\hat{\mu} - \mu_0 = O(n^{-1/2})$, 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 (5), can be interpreted as the c.d.f. of $\hat{\tau}$ under μ. In order to describe the *joint* distribution of $\hat{\tau}$ and γ, we also need the distribution of γ conditional on $\hat{\tau}$. Let us denote by $g(\gamma \mid \tau)$ the density of γ conditional on $\hat{\tau} = \tau$ under the DGP μ_0. Since $g(\gamma \mid \tau)$ is a density,

$$\int_{-\infty}^{\infty} g(\gamma \mid \tau)\, d\gamma = 1 \quad \text{for all } \tau \in [0, 1]. \tag{14}$$

With this specification, we can compute the true size of the bootstrap test as the probability under μ_0 that $\hat{\tau} \leq Q(\alpha, \hat{\mu})$. By (13), this true size is

$$\int_{-\infty}^{\infty} d\gamma \int_0^{Q+n^{-k/2}\sigma_\gamma\gamma} dS(\tau)\, g(\gamma \mid \tau), \tag{15}$$

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 (15) can be split into two parts, as follows:

$$\begin{aligned} &\int_0^Q dS(\tau) \int_{-\infty}^{\infty} d\gamma\, g(\gamma \mid \tau) \\ &\quad + \int_{-\infty}^{\infty} d\gamma \int_0^{n^{-k/2}\sigma_\gamma\gamma} d\tau\; S'(Q+\tau)\, g(\gamma \mid Q+\tau), \end{aligned} \tag{16}$$

where S' is the derivative of $S(\tau)$. Because of (14), the integral over γ in the first term of (16) 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 (16), which is clearly of order no more than $n^{-k/2}$.

This last point is the result of Beran (1988) for asymptotically pivotal test statistics. However, we have gone beyond his analysis by giving an explicit expression for the order $n^{-k/2}$ size discrepancy. Our subsequent results will follow from an examination of this explicit expression.

The P value discrepancy of the bootstrap test at nominal size α is the second term in (16), which, if g is smooth enough, we may write as

$$n^{-k/2}\sigma_\gamma \int_{-\infty}^{\infty} d\gamma\, \gamma\, \big(g(\gamma \mid \alpha) + O(n^{-k/2})\big)\big(1 + O(n^{-l/2})\big), \tag{17}$$

since, by (10), $Q \equiv Q(\alpha, \mu_0) = \alpha + O(n^{-l/2})$, and this, along with (6), gives $S'(Q) = 1 + O(n^{-l/2})$.

If we consider only the leading-order term in (17), we see that it has a simple interpretation. It is the expectation, conditional on $\hat{\tau} = \alpha$, of $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$. Thus it is the bias, conditional on $\hat{\tau} = \alpha$, of the bootstrap estimate of the size-α critical value. 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 $O(n^{-(k+1)/2})$ or lower, the term in (17) of order $n^{-k/2}$ 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 on average be balanced to leading order by those $\hat{\mu}$ which underestimate it. Specifically, if

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

for $i \leq k$, the P value discrepancy (17) is of order $n^{-(k+i)/2}$.

The simplicity of the interpretation of the leading-order term in (17) arises from the fact that we expressed the test statistic in approximate P value form. Any bias in the bootstrap estimate of the true quantile of such a P value is thus by construction the P value discrepancy to leading order — to leading order only, because the test statistic is only an approximate P value. It is natural as well that what is needed is the bias conditional on the test statistic taking on a value at the margin of rejection at level α. A bias of the bootstrap quantile away from this margin will have no effect on the rejection probability of the test at this level.

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 hypothesis testing terms, the critical values for a two-tailed test based on such a statistic computed from the true DGP and from the bootstrap DGP differ only at that order; in other words, $k = 3$. (This result is plainly unaffected by our use of statistics in approximate P value form rather than approximately standard normal form.) Hall goes on to show that the coverage error of a bootstrap confidence interval is of still lower order, namely, n^{-2} in general. This translates as a P value discrepancy of that order. Examination of Hall's derivation of this result reveals that the additional refinement is due, as (17) 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.

It is in fact quite easy to see why, in this case of a symmetric two-tailed test, what appears to be the leading-order term in (17), namely, the expectation of γ conditional on $\hat{\tau} = \alpha$, vanishes at the leading order. The condition that the approximate P value $\hat{\tau}$ take on the value α corresponds, in terms of the asymptotically standard normal statistic, to one of two possibilities: The latter statistic can be equal either to the higher, positive, critical point, or to its negative. Under the null, because the test is symmetric, both of these events are equally probable, at least 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, will be equal and opposite in sign. Thus the expectation of γ conditional on $\hat{\tau} = \alpha$ vanishes to leading order.

In general, without any requirement of a symmetric two-sided test, the hypothesis testing context enables us to strengthen (17) if a further condition is satisfied. This condition is that γ and $\hat{\tau}$ should be asymptotically independent. We show in a moment that this condition leads to a further asymptotic refinement.

The asymptotic independence of γ and $\hat{\tau}$ 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. This last fact can be illustrated most simply in the context of a linear regression model. Consider the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ in the regression

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u},$$

where $\boldsymbol{X}_1$ is $n \times k_1$ and $\boldsymbol{X}_2$ is $n \times k_2$. The F statistic for this hypothesis is (see, for instance, Davidson and MacKinnon (1993), Chapter 3):

$$\frac{n-k}{r}\,\frac{\|\boldsymbol{P}_{M_1X_2}\boldsymbol{M}_1\boldsymbol{y}\|^2}{\|\boldsymbol{M}_{M_1X_2}\boldsymbol{M}_1\boldsymbol{y}\|^2},$$

where $\boldsymbol{M}_1 \equiv \mathbf{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ is the orthogonal projection on to the orthogonal complement of the span of the k_1 columns of $\boldsymbol{X}_1$, and $\boldsymbol{P}_{M_1X_2}$ and $\boldsymbol{M}_{M_1X_2}$ respectively project orthogonally on to and off the span of the columns of $\boldsymbol{M}_1\boldsymbol{X}_2$. Thus the F statistic depends only on the components of the error vector $\boldsymbol{u}$ that are orthogonal to the columns of $\boldsymbol{X}_1$. On the other hand, the error in the OLS estimate of $\boldsymbol{\beta}_1$ under the null is $(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{u}$, which depends only on the components of $\boldsymbol{u}$ that are in the span of the columns of $\boldsymbol{X}_1$. Under normality, this implies the independence of the estimate of $\boldsymbol{\beta}_1$ under the null and the F statistic. Without normality, it implies their asymptotic independence.

More generally, if $\hat{\boldsymbol{\theta}}$ is an extremum estimator that satisfies first-order conditions in the interior of the parameter space of the null hypothesis, the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ 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.

This means that, for the parametric bootstrap, the condition of asymptotic independence of $\hat{\tau}$ 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 $\hat{\tau}$. This argument clearly also applies to cases in which only a conditional model for the dependent variable is fully parametric.

Asymptotic independence can often be achieved with little trouble for the nonparametric 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 function of the residuals. 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 of the error terms. To leading order asymptotically, it will have the form

$$n^{-1/2}\sum_{t=1}^{n} x_t u_t, \quad \text{where } \sum_{t=1}^{n} x_t = 0 \text{ and } \sum_{t=1}^{n} x_t^2 = 1. \tag{18}$$

But all we need is asymptotic independence of the statistic and the quantiles of the bootstrap distribution. For linear regression models with a constant term, Hall (1992, section 4.3.3) shows that the leading-order random term in the bootstrap quantiles of such a t statistic is proportional to the leading-order random term in the skewness of the residuals, which can be expressed as an expression of the form

$$n^{-1/2}\sum_{t=1}^{n} f(u_t), \quad \text{with } E\big(f(u_t)\big) = 0. \tag{19}$$

It is immediate that (18) and (19) are both asymptotically normal with zero mean, and that their covariance is zero. They are therefore asymptotically independent, as desired. This type of result undoubtedly holds more generally for the nonparametric bootstrap applied to regression models. In addition, as this example shows, in the frequently encountered case of asymptotic normality, all that is needed is a zero asymptotic covariance.

Let us now demonstrate that, if we assume that γ and $\hat{\tau}$ are asymptotically independent, there is an asymptotic refinement in the P value discrepancy. Under that assumption, we may write

$$g(\gamma \mid \tau) = g(\gamma)\big(1 + n^{-j/2} f(\gamma, \tau)\big), \tag{20}$$

where $g(\gamma)$ is the asymptotic marginal distribution of γ, $j \geq 1$ is a suitable integer, and $f(\gamma, \tau)$ is of order unity as $n \to \infty$.

For any valid bootstrap procedure, $\hat{\mu}$ must be a consistent estimator of μ_0. Therefore, $Q(\alpha, \hat{\mu})$ must be a consistent estimator of $Q(\alpha, \mu_0)$, and we have that

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

for otherwise $Q(\alpha, \hat{\mu})$ would be biased at leading order and thus not consistent. Thus (17) becomes

$$n^{-(k+j)/2}\sigma_\gamma \int_{-\infty}^{\infty} d\gamma\ \gamma\, g(\gamma)\, f(\gamma, \alpha) + O(n^{-(k+j+1)/2}). \tag{21}$$

The interpretation of (21) is the same as that of (17). The first term is the bias, conditional on $\hat{\tau} = \alpha$, of the size-α critical value based on the bootstrap DGP $\hat{\mu}$. When $k = 2$ 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 first determinant of size distortion is the order of the discrepancy between the bootstrap critical value and the true critical value. In (13), we specified that it was $O(n^{-k/2})$. In most cases, with root-n consistency of parameter estimates for the parametric bootstrap, or with root-n consistency of the nonparametric bootstrap distribution, $k = l + 1$ with $l \geq 1$. Provided an asymptotically pivotal statistic is used, $l = 1$ and $k = 2$ in the worst case. In (17), we see that the bootstrap P value will be incorrect only at $O(n^{-k/2})$ at most, in accord with the standard result cited at the end of the last section.

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 $O(n^{-(k+i)/2})$, where $n^{-i/2}$ is the order of the bias of the bootstrap critical value. Secondly, when the discrepancy is asymptotically independent of the test statistic $\hat{\tau}$, the bootstrap P value will be incorrect only at $O(n^{-(k+j)/2})$, where $j/2$ is the highest order at which the conditional distribution of the bootstrap critical value differs from its asymptotic marginal distribution. This second case always holds for the parametric bootstrap if it is based on a regular extremum estimator under the null, and it also holds for many nonparametric bootstrap procedures in regression models. In cases with no specific sources of refinements, 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 $O(n^{-1/2})$. By focusing on the bias of the bootstrap critical value for a statistic in approximately P value form, we have thus provided a unifying framework for understanding a variety of seemingly unrelated results in the statistical literature.

4. 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,

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

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 τ is asymptotically pivotal and that τ and

e are asymptotically independent. Then the joint density of τ and e can be written as

$$S'(\tau, \theta_0)\phi(e)\bigl(1 + n^{-1/2}a(e, \tau)\bigr). \tag{22}$$

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 $a(e, \tau) = O(1)$. In principle, we could have $n^{-j/2}$ instead of $n^{-1/2}$ in (22), as we did in (20). 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 (21) 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

$$-n^{-(l+2)/2}\left(s_\theta(\alpha, \theta)\int_{-\infty}^{\infty} e\,\phi(e)\,a(e, \alpha)\,de + \frac{1}{2}s_{\theta\theta}(\alpha, \theta)\right) + O(n^{-(l+3)/2}), \tag{23}$$

where s_θ and $s_{\theta\theta}$ denote the first and second derivatives of $s(\alpha, \theta)$ with respect to θ. For given α, (23) is simply a function of θ.

There are two leading order terms in expression (23), and these are of order at most $n^{-3/2}$, as we would expect from (21). 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 (23) 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 (23). 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 (22). 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 (23) 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 (23) 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 (21), 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 (23) 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 (23) applies only to the one-parameter case, these results must evidently be true more generally.

5. Bootstrap *J* Tests

In this section, we present some simulation results which are designed to see whether the theoretical results of Sections 3 and 4 are useful in practice. They concern the J test of nonnested hypotheses. 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, and we are not the first to suggest doing so; see Fan and Li (1995) and Godfrey (1996). However, our simulations provide much more precise results than those in earlier papers, and they do so in the context of the analysis of Sections 3 and 4.

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

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

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 $\boldsymbol{\gamma}$ is $k_2 \times 1$. The J test statistic is the ordinary t statistic for $\alpha = 0$ in the artificial regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \alpha \boldsymbol{P}_Z \boldsymbol{y} + \text{residuals}, \tag{24}$$

where $\boldsymbol{P}_Z = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$. Thus $\boldsymbol{P}_Z\boldsymbol{y}$ 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, which as before we denote by $\hat{\tau}$, by running regression (24) after obtaining the fitted values from the H_2 model. Then we draw B sets of bootstrap error terms $\boldsymbol{u}_j^*$ and use them along with the parameter estimates $\hat{\boldsymbol{\beta}}$ from H_1 to generate B bootstrap samples $\boldsymbol{y}_j^*$, $j = 1, \ldots, B$, according to the equation

$$\boldsymbol{y}_j^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{u}_j^*.$$

There are numerous ways in which the error terms $\boldsymbol{u}_j^*$ can be drawn, four of which will be described below. Using each of these bootstrap samples, we calculate a test statistic τ_j^*, and we then compute the estimated bootstrap P value as

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

where $I(\cdot)$ is an indicator function, equal to 1 if its argument is true and equal to zero otherwise. The absolute values are needed here because the J test is a two-tailed test.

We consider four different ways of generating the $\boldsymbol{u}_j^*$. For the parametric bootstrap, which we will call b_0, they are simply n-vectors of independent draws from the $N(0, s^2)$ distribution, where s is the OLS estimate of σ from estimating H_1. For the simplest nonparametric bootstrap, which we will call b_1, they are obtained by resampling with replacement from the vector of residuals $\hat{u}_t$ from H_1. A slightly more complicated form of nonparametric bootstrap, which we will call b_2, generates the $\boldsymbol{u}_j^*$ by resampling with replacement from the vector with typical element

$$\big(n/(n - k_1)\big)^{1/2}\hat{u}_t.$$

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 $\boldsymbol{u}_j^*$ 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 $\boldsymbol{P}_X \equiv \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. 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.

If $\acute{s}$ denotes the estimated standard error from regression (24), the J test statistic can be written as

$$\frac{\boldsymbol{y}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{y}}{\acute{s}(\boldsymbol{y}^\top \boldsymbol{P}_Z \boldsymbol{M}_X \boldsymbol{P}_Z \boldsymbol{y})^{1/2}}, \tag{25}$$

where $\boldsymbol{M}_X \equiv \mathbf{I} - \boldsymbol{P}_X$. It is straightforward to show that, under H_1, the statistic (25) 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 the experiments we report. All elements of $\boldsymbol{\beta}^*$ were equal.

Figure 3 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 3, 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 4 may therefore come as a surprise. This figure shows the proportion of replications with P values less than .05, as a function of the sample size n, for four values of θ and 21 different sample sizes: 8, 9, 10, 11, 12, 14, 16, 18, 20, and then all multiples of 5 up to 80. Each experiment used 100,000 replications, and there were $B = 399$ bootstrap samples for each replication. The former number may seem rather large, but, as we shall see, the bootstrap test works extremely well. Thus, in order to detect any pattern in the results, it was 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, partly because most of the experiments involve very small sample sizes, and partly because we are computing tail-area probabilities; see Davidson and MacKinnon (1992).

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 3 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, since 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 5.38% and 5.37% of the time. In comparison, the t test rejects 37.91% of the time.

One reason for the remarkably good performance of the J test is that θ is estimated quite precisely. By using the delta method, it is easy to show that the asymptotic variance of $\hat{\theta}$ for the experimental design we used is $\frac{1}{n}(1 + \frac{1}{2}\theta^2)$. From this formula, it is apparent that, except for very small values of n and/or very large values of θ, the standard error of $\hat{\theta}$ is always substantially less than 1. This means that the curvature of the P value functions which is evident in Figure 3 has only a modest effect on the performance of the bootstrap, even when θ is quite close to 0.

These results appear to provide strong support for the theory of Sections 3 and 4. 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, even though the results of Fan and Li (1995) and Godfrey (1996) also provide no evidence to contradict such a claim. There undoubtedly exist situations in which PVFs are even steeper or more sharply curved than the ones in Figure 3, or in which the parameters on which the bootstrap distribution depends are estimated less precisely, 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, such as $\theta = 0$ in Figure 3.

The result that the bootstrap J test works extremely well, even in very small samples, is consistent with the results of Rayner (1990), who studied bootstrap tests of the slope coefficient in an AR(1) model with a constant term. We also obtained extremely good results, not reported here, when we studied a bootstrap test for AR(1) errors in a regression model with a lagged dependent variable. Thus there seems to be good reason to believe that bootstrap tests may often work extremely well, at least for tests of regression models in regression directions. However, the results of Horowitz (1994), for the information matrix test in a probit model, suggest that bootstrap tests may not always work quite as well in other cases.

6. 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, and can indeed make use of existing results on 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 using the parametric bootstrap, 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 quantity which determines the order of the discrepancy between bootstrap and true critical values, denoted by k in Section 3, is often greater than 2 in testing situations that arise frequently in econometrics. The case of two-sided tests based on asymptotically standard normal statistics, or of asymptotically chi-squared statistics, has already been mentioned. In addition, in the context of linear regression models with constant terms, Hall (1992) shows that there is a refinement associated with bootstrap inference on the slope coefficients, one which, taken in conjunction with that afforded by two-sided tests, can lead to a value of 3 for k in many circumstances. It seems likely that the excellent performance of the bootstrapped J test described in Section 5 is due to such specific additional refinements.

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 4, 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 example studied in detail in Section 5. Thus the analysis of the determinants of size distortion 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 4. 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 $\hat{\tau}$ 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 $\hat{\tau}$ 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.

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 seem to be 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 might 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. Of course, we cannot claim that bootstrap tests will always perform this well, especially for models that do not fit into the regression framework.

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 (23) 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 (13).

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 (20) 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 (20).

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 (21), 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 $\tau = \alpha$. The density of e conditional on τ is just the product of the last two factors in (22). Thus, since e and γ are related, without reference to τ, by (A.04) and (A.05), the density of γ conditional on $\tau = \alpha$ 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

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

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

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

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 (20) 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 (21), 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

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

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:

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

Expression (23) 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.

Beran, R. (1986). "Simulated power functions," *Annals of Statistics*, 14, 151–173.

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, 1305–1329.

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.

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

Fan, Y., and Q. Li (1995). "Bootstrapping J-type tests for non-nested regression models," *Economics Letters*, 48, 107-112.

Godfrey, L. G. (1996). "Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap," University of York, mimeo.

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 J. L. Horowitz (1996). "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrica*, 64, 891–916.

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.

Hinkley, D. V. (1988). "Bootstrap methods," *Journal of the Royal Statistical Society*, Series B. 50, 321–337.

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

Rayner, R. K. (1990). "Bootstrapping *p* values and power in the first-order autoregression: a Monte Carlo investigation," *Journal of Business and Economic Statistics*, 8, 251–263.

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.

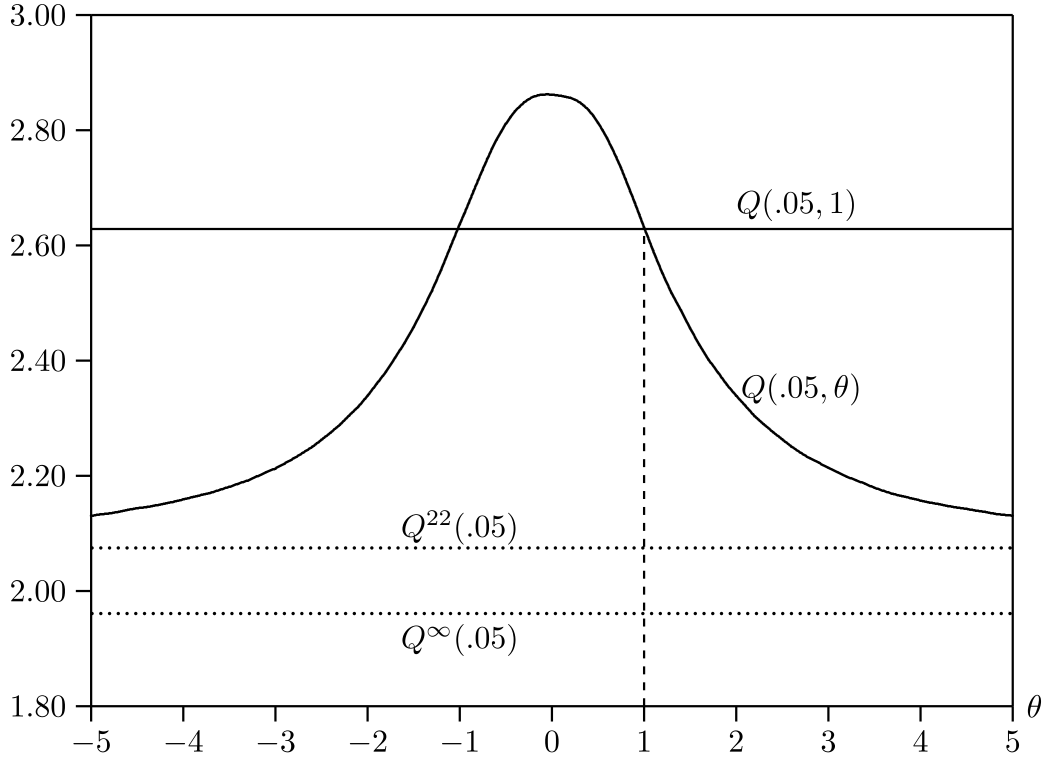

Figure 1. A Critical Value Function

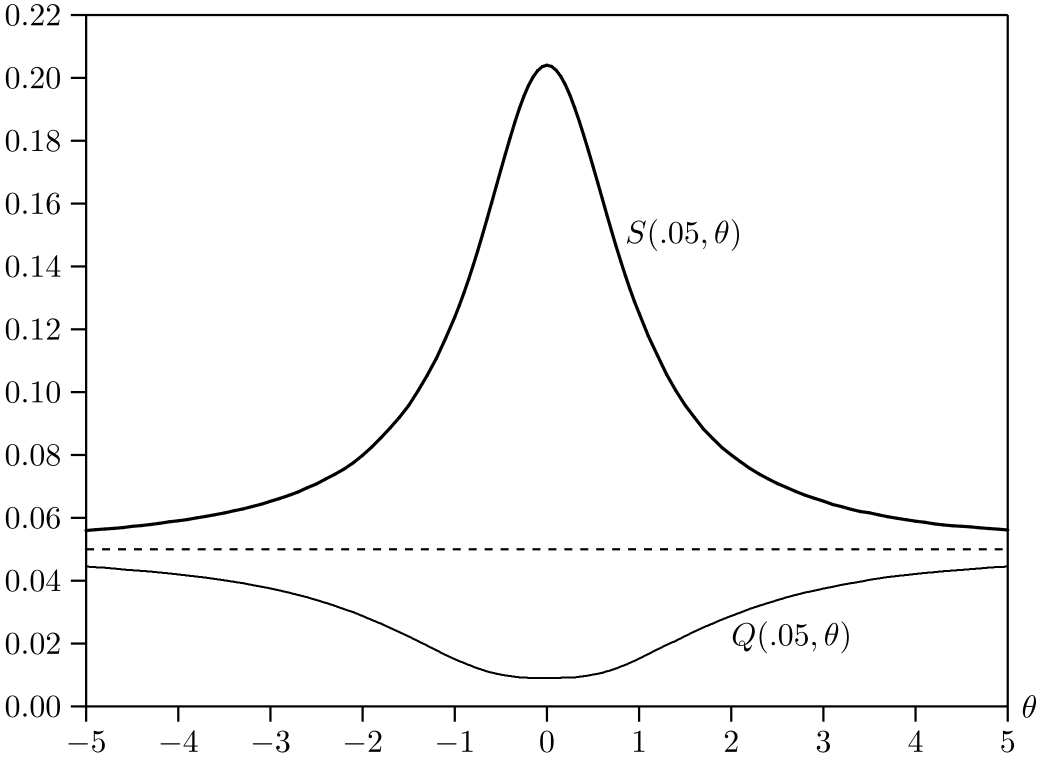

Figure 2. A *P* Value Function and its Inverse

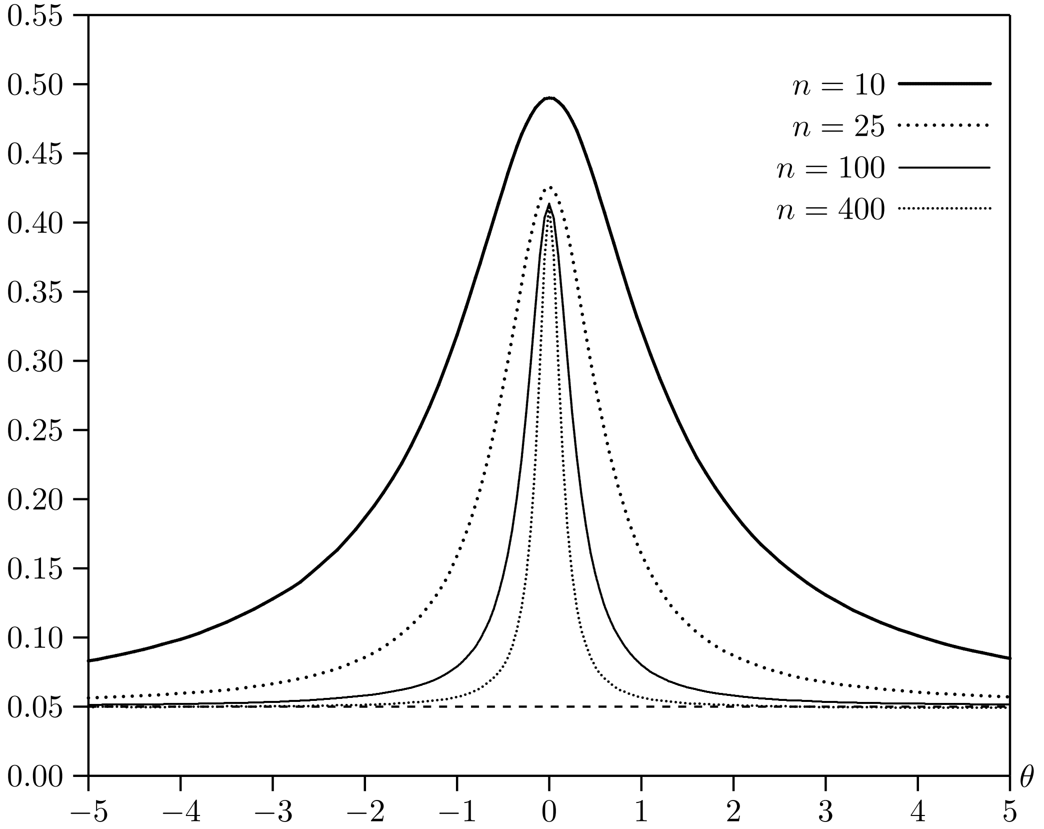

Figure 3. ***P*** **Value Functions for** ***J*** **Tests**

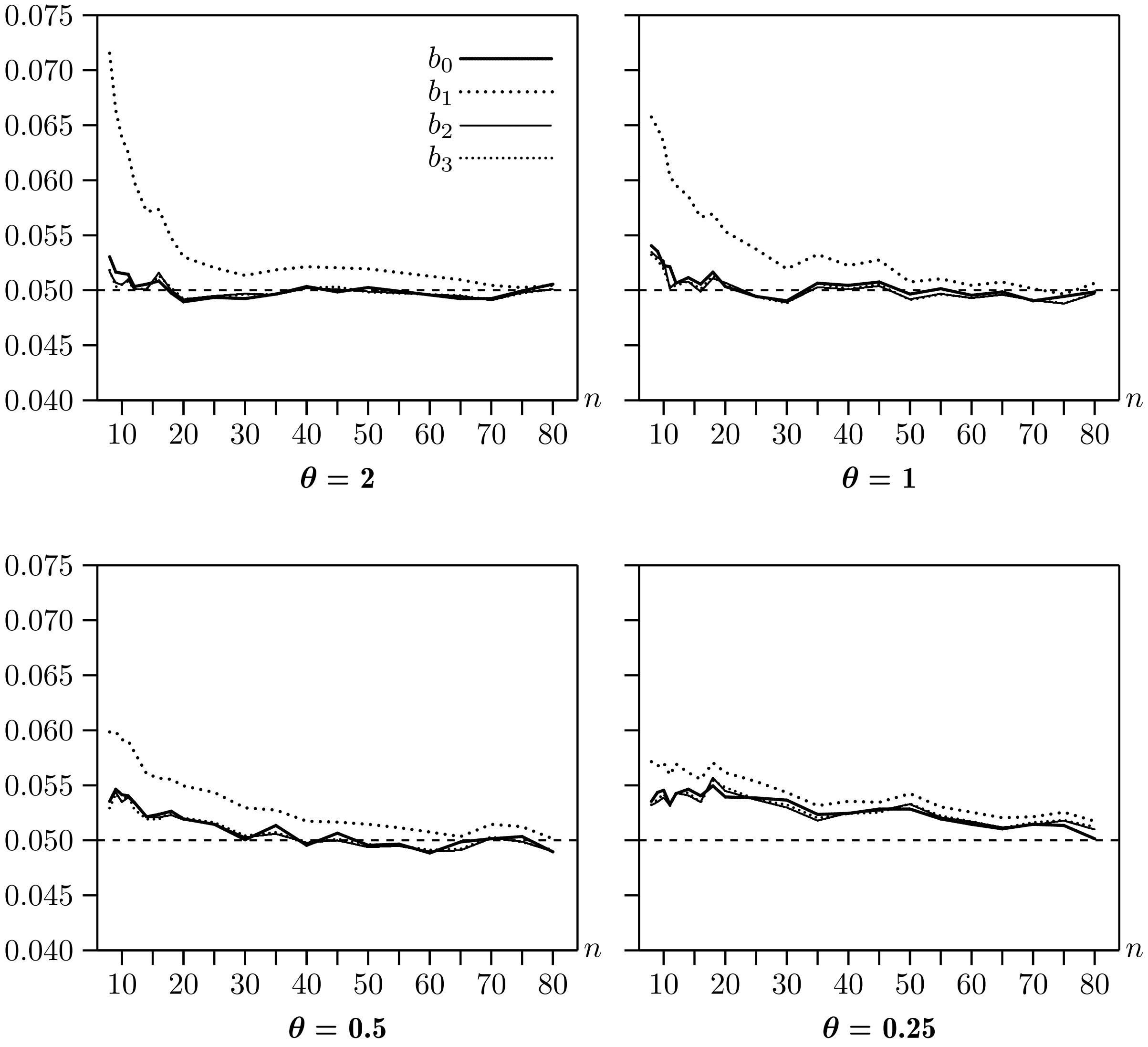

Figure 4. Estimated P Values for Bootstrap J Tests at .05 Level