BOOTSTRAP TESTS: HOW MANY BOOTSTRAPS?

Russell Davidson

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

and

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

James G. MacKinnon

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

Key Words and Phrases: bootstrap test; test power; pretest.

JEL Classification: C12, C15.

ABSTRACT

In practice, bootstrap tests must use a finite number of bootstrap samples. This means that the outcome of the test will depend on the sequence of random numbers used to generate the bootstrap samples, and it necessarily results in some loss of power. We examine the extent of this power loss and propose a simple pretest procedure for choosing the number of bootstrap samples so as to minimize experimental randomness. Simulation experiments suggest that this procedure will work very well in practice.

1. INTRODUCTION

As a result of remarkable increases in the speed of digital computers, the bootstrap has become increasingly popular for performing hypothesis tests. In econometrics, the use of the bootstrap for this purpose has been advocated by Horowitz (1994), Hall and Horowitz (1996), Li and Maddala (1996), and others. Although there are many ways to use the bootstrap for hypothesis testing, in this paper we emphasize its use to compute P values. While other ways of performing bootstrap tests are fundamentally equivalent, the P value approach is the simplest to analyze.

Following the P value approach, one first computes a test statistic, say $\hat{\tau}$, in the usual way, and estimates whatever parameters are needed to obtain a data generating process (DGP) that satisfies the null hypothesis. The distribution of the random variable τ of which $\hat{\tau}$ is a realization under this "bootstrap DGP" serves to define the theoretical or ideal bootstrap P value, $p^*(\hat{\tau})$, which is just the probability that $\tau > \hat{\tau}$ under the bootstrap DGP. Normally this probability cannot be calculated analytically, and so it is estimated by simulation, as follows. One draws B bootstrap samples from the bootstrap DGP, each of which is used to compute a bootstrap test statistic τ_j^* in exactly the same way as the real sample was used to compute $\hat{\tau}$. For a one-tailed test with a rejection region in the upper tail, the bootstrap P value may then be estimated by the proportion of bootstrap samples that yield a statistic greater than $\hat{\tau}$:

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

where $I(\cdot)$ is the indicator function. As $B \to \infty$, it is clear that the estimated bootstrap P value $\hat{p}^*(\hat{\tau})$ will tend to the ideal bootstrap P value $p^*(\hat{\tau})$.

There are two types of error associated with bootstrap testing. The first may occur whenever a test statistic is not pivotal. Although most test statistics used in econometrics are *asymptotically* pivotal, usually with known asymptotic distributions, most of them are not pivotal in finite samples. This means that the distribution of the statistic depends on unknown parameters or other unknown features of the DGP. As a consequence, bootstrap P values will generally be somewhat inaccurate, because of the differences between the bootstrap DGP and the true DGP. Nevertheless, inferences based on the bootstrap applied to asymptotically pivotal statistics will generally be more accurate than inferences based on asymptotic theory, in the sense that the errors are of lower order in the sample size; see Beran (1988), Hall (1992), and Davidson and MacKinnon (1999).

The second type of error, which is the subject of this note, arises because B is necessarily finite. An ideal bootstrap test rejects the null hypothesis at level α whenever $p^*(\hat{\tau}) < \alpha$. A feasible bootstrap test rejects it

whenever $\hat{p}^*(\hat{\tau}) < \alpha$. If drawing the bootstrap samples and computing the τ_j^* were sufficiently cheap, we would choose B to be extremely large, thus ensuring that the ideal and feasible tests almost always led to the same outcome. In practice, however, computing the τ_j^* is sometimes not particularly cheap. In such cases, we want B to be fairly small.

There are two undesirable consequences of using a finite number of bootstrap samples. The first is simply that the outcome of a test may depend on the sequence of random numbers used to generate the bootstrap samples. The second is that, whenever $B < \infty$, there will be some loss of power, as discussed in Hall and Titterington (1989), among others. This loss of power is often small, but, as we will see in Section 2, it can be fairly large in some cases. The principal contribution of this paper, which we introduce in Section 3, is a method for choosing B, based on pretesting, designed so that, although B is fairly small on average, the feasible and ideal tests rarely lead to different outcomes. In Section 4, we assess the performance of this procedure by simulation methods. In Section 5, we examine the performance of an alternative procedure recently proposed by Andrews and Buchinsky (1998).

2. POWER LOSS FOR FEASIBLE BOOTSTRAP TESTS

The power loss associated with using finite B has been investigated in the literature on Monte Carlo tests, which are similar to bootstrap tests but apply only to pivotal test statistics. When the underlying test statistic is pivotal, a bootstrap test is equivalent to a Monte Carlo test. The idea of Monte Carlo tests is generally attributed to Dwass (1957) and Barnard (1963), and early papers include Hope (1968) and Marriott (1979). A recent application of Monte Carlo tests in econometrics can be found in Dufour and Kiviet (1998). One key feature of Monte Carlo tests is that B must be chosen so that $\alpha(B+1)$ is an integer if the test is to be exact; see the Dufour and Kiviet paper for details. Therefore, for $\alpha = .05$, the smallest possible value of B for an exact test is 19, and for $\alpha = .01$, the smallest possible value is 99. Although it is not absolutely essential to choose B in this way for bootstrap tests when the underlying test statistic is nonpivotal, since they will not be exact anyway, it is certainly sensible to do so.

Using a finite value of B inevitably results in some loss of power, because the test has to allow for the randomness in the bootstrap samples. The issue of power loss in Monte Carlo tests was first investigated for a rather special case by Hope (1968). Subsequently, Jöckel (1986) obtained some fundamental theoretical results for a fairly wide class of Monte Carlo tests, and his results are immediately applicable to bootstrap tests.

For any pivotal test statistic and any fixed DGP, we can define the "size-power" function, $\eta(\alpha)$, as the probability under that DGP that the test

will reject the null when the rejection probability under the null is α. Since the statistic is pivotal, α is well defined. This function is precisely what we plot as a size-power curve using simulation results; see, for example, Davidson and MacKinnon (1998). It is always true that $\eta(0) = 0$ and $\eta(1) = 1$, and we need $\eta(\alpha) > \alpha$ for $0 < \alpha < 1$ for the test to be consistent. Thus, in general, we may expect the size-power function to be concave. For tests that follow standard noncentral distributions, such as the noncentral χ^2 and the noncentral F, this is indeed so. However, it is not generally so for every test and for every DGP that might have generated the data.

Let the size-power function for a bootstrap test based on B bootstrap samples be denoted by $\eta^*(\alpha, m)$, where $\alpha(B + 1) = m$. When the original test statistic is pivotal, $\eta(\alpha) \equiv \eta^*(\alpha, \infty)$. Under this condition, Jöckel (1986) proves the following two results:

(i) If $\eta(\alpha)$ is concave, then so is $\eta^*(\alpha, m)$.

(ii) Assuming that $\eta(\alpha)$ is concave, $\eta^*(\alpha, m + 1) > \eta^*(\alpha, m)$.

Therefore, increasing the number of bootstrap samples will always increase the power of the test. Just how much it will do so depends in a fairly complicated way on the shape of the size-power function $\eta(\alpha)$. Pivotalness is not needed if we wish to compare the powers of an ideal and feasible bootstrap test. We simply have to interpret $\eta(\alpha)$ as the size-power function for the ideal bootstrap test. Provided this function is concave, Jöckel's two results apply. Thus we conclude that the feasible bootstrap test will be less powerful than the ideal bootstrap test whenever the size-power function for the ideal test is concave.

Jöckel presents a lower bound on the ratio of $\eta^*(\alpha, m)$ to $\eta(\alpha)$, which suggests that power loss can be quite substantial, especially for small rejection probabilities under the null. From expression (4.19) of Davison and Hinkley (1997), which provides a simplified version of Jöckel's bound,[1] we find that

$$\eta(\alpha) - \eta^*(\alpha, m + 1) \leq \eta(\alpha)\left(\frac{1 - \alpha}{2\pi m}\right)^{1/2}. \tag{2}$$

Thus the maximum power loss increases with $\eta(\alpha)$, may either rise or fall with α, and is proportional to the square root of $1/(B + 1)$.

It is interesting to study how tight the bound (2) is in a simple case. We accordingly conducted a small simulation experiment in which we generated t statistics for the null hypothesis that $\gamma = 0$ in the model

$$y_t = \gamma + u_t, \quad u_t \sim N(0, 1), \quad t = 1, \ldots, 4. \tag{3}$$

These t statistics were then converted to P values. For the ideal bootstrap, this was done by using the CDF of the $t(3)$ distribution so as to obtain p^*.

[1] Note that the inequality goes the wrong way in the first printing of the book.

For the feasible bootstrap, it was done by drawing bootstrap test statistics from the $t(3)$ distribution and using equation (1), slightly modified to allow for the fact that this is a two-tailed test, to obtain $\hat{p}^*$. There were one million replications.

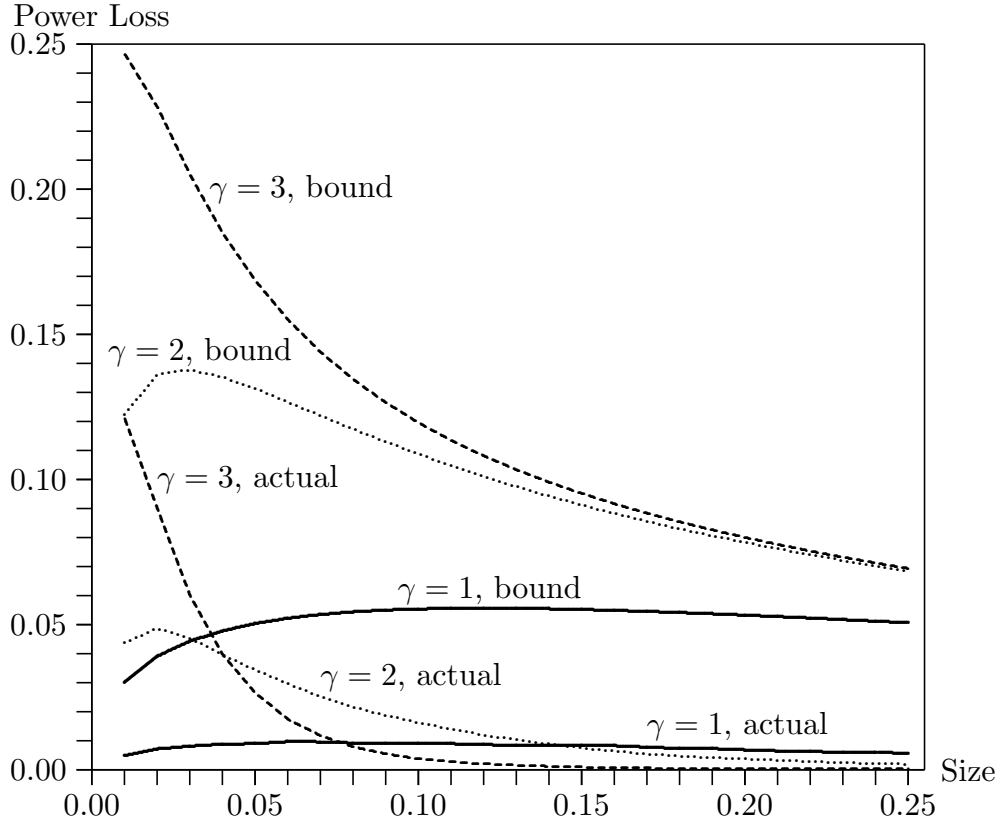

FIG. 1 Power Loss from Bootstrapping, $B = 99$ (Bound and Actual)

From Jöckel's results, we know that it is only the size-power function that affects power loss. The details of (3) (in particular, the very small sample size) were chosen solely to keep the cost of the experiments down. They affect the results only to the extent that they affect the shapes of the size-power curves. We performed experiments for $\gamma = 0.5, 1.0, \ldots, 3.0$ and present results for $\gamma = 1.0$, 2.0, and 3.0, since these three cases seem to be representative of tests with low, moderate, and high power. When $\gamma = 1.0$, the test rejected 7.6% and 29.0% of the time at the .01 and .05 levels; when $\gamma = 2.0$, it rejected 30.9% and 75.7% of the time; and when $\gamma = 3.0$, it rejected 62.0% and 96.7% of the time.

Figure 1 shows the actual power loss observed in our experiments, together with the loss implied by the bound (2), for $B = 99$, which is the

smallest value of B that is commonly suggested. The shape of the bounding power loss function is quite similar to the shape of the actual power loss function, but the actual power loss is always considerably smaller than the bound. The largest loss occurs for small values of α when the test is quite powerful. In the worst case, when $\gamma = 3.0$ and $\alpha = .01$, this loss is substantial: The ideal bootstrap test rejects 62% of the time, and the feasible bootstrap test rejects only 50% of the time.

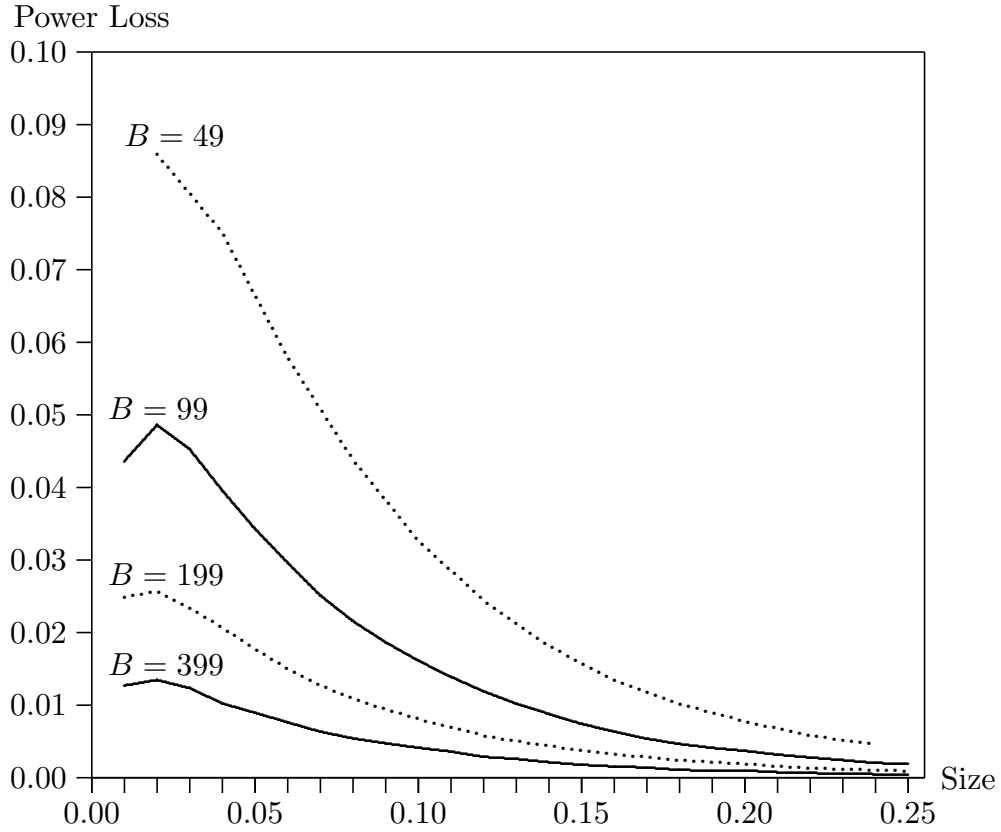

FIG. 2 Power Loss from Bootstrapping, $\gamma = 2$

Another interesting finding is shown in Figure 2. The bound (2) implies that power loss will be proportional to $(B+1)^{-1/2}$, but, as the figure makes clear, our results appear to show that it is actually proportional to $(B+1)^{-1}$. This suggests that, in regular cases like the one studied in our experiments, the bound (2) becomes increasingly conservative as B becomes larger.

In order to avoid a power loss of more than, say, 1%, it is necessary to use a rather large number of bootstrap samples. If our simulation results can be relied upon, $B = 399$ would seem to be about the minimum for a test at the .05 level, and $B = 1499$ for a test at the .01 level. If there are cases in which the bound (2) is tight, then very much larger values of B are needed.

3. CHOOSING B BY PRETESTING

Up to this point, we have assumed that B is chosen in advance. However, as Andrews and Buchinsky (1988) point out, if one wishes to bound the proportional error of a feasible bootstrap P value, the minimum number of bootstraps needed depends on the ideal bootstrap P value. They develop an algorithm for a data-dependent choice of B, which we look at more closely in Section 5, designed to control the proportional error of a bootstrap P value in all cases. Nevertheless, although P values are more informative than the yes/no result of a test at a given level α, it is often the case that specific values of α, like .05 or .01, are of special interest. In this section, we propose for such cases a simple pretesting procedure for determining B endogenously.

Issues other than test power affect any such choice. When B is finite, the randomness of a bootstrap P value, or the result of a bootstrap test at level α, comes from two sources, the data and the simulations, these two sources being independent. We wish the outcome of a test to depend only on the first source. One way to achieve that would be to condition on the simulation randomness, but that would impose the constraint of using the same random number generator with the same seed for all bootstrap inference. Otherwise, all one can do is to seek to minimize the effect of simulation randomness. Another issue is, of course, computing time. *Ceteris paribus*, it makes sense to minimize expected computing time, where the expectation is with respect to both sources of randomness. Our problem is to do so without inducing undue power loss or an unacceptably high probability of a test result in conflict with that based on the ideal bootstrap P value.

A couple of examples serve to illustrate these issues when some level α is of particular interest. If $\hat{\tau} > \tau_j^*$ for every one of B bootstrap samples, B need not be large for us to conclude that we should reject the null hypothesis. The probability of this event occurring by chance if $p^*(\hat{\tau})$ is equal to α is $(1-\alpha)^B$. For $B = 99$ and $\alpha = .05$, this probability is .006. Thus, if $p^*(\hat{\tau})$ is greater than or equal to .05, .006 is an upper bound on the probability that $\hat{\tau} > \tau_j^*$ for each of 99 bootstrap samples. Similarly, suppose that $\hat{p}^*(\hat{\tau})$ based on $B = 99$ is substantially greater than .05. According to the binomial distribution, the probability that $\tau_j^* > \hat{\tau}$ 11 or more times out of 99 if $p^*(\hat{\tau}) = .05$ is .004. Thus, if in fact $p^*(\hat{\tau}) \leq .05$, it is highly improbable that 11 or more out of 99 bootstrap samples will produce test statistics more extreme than $\hat{\tau}$.

These examples suggest a pretesting procedure in which we start with a relatively small value of B and then increase it, if necessary, until we are confident, at some prechosen significance level, that $p^*(\hat{\tau})$ is either greater or less than α. If the procedure stops with a small value of B, $\hat{p}^*(\hat{\tau})$ may differ substantially from $p^*(\hat{\tau})$, but only when $p^*(\hat{\tau})$ is not close to α, thus ensuring low probability that the feasible and ideal bootstrap tests yield

different outcomes.

To implement the procedure, we must choose β, the level of the pretest (say .001), and two rather arbitrary parameters that can be expected to have little impact on the result of the procedure: $B_{\min}$, the initial number of bootstrap samples (say, 99), and $B_{\max}$, the maximum number of bootstrap samples (say, 12,799). The second parameter effectively bounds computing time, and avoids the problem that, if $p^*(\hat{\tau})$ happens to be very close to α, then a huge number of bootstraps would be needed to determine whether it is greater or smaller than α. The procedure can be set out as follows:

1. Calculate $\hat{\tau}$, set $B = B_{\min}$ and $B' = B_{\min}$, and calculate τ_j^* for $B = B_{\min}$ bootstrap samples.
2. Compute $\hat{p}^*(\hat{\tau})$ based on B bootstrap samples. Depending on whether $\hat{p}^*(\hat{\tau}) < \alpha$ or $\hat{p}^*(\hat{\tau}) > \alpha$, test either the hypothesis that $p^*(\hat{\tau}) \geq \alpha$ or the hypothesis that $p^*(\hat{\tau}) \leq \alpha$ at level β. This may be done using the binomial distribution or, if αB is not too small, the normal approximation to it. If $\hat{p}^*(\hat{\tau}) < \alpha$ and the hypothesis that $p^*(\hat{\tau}) \geq \alpha$ is rejected, or if $\hat{p}^*(\hat{\tau}) > \alpha$ and the hypothesis that $p^*(\hat{\tau}) \leq \alpha$ is rejected, stop.
3. If the algorithm gets to this step, set $B = 2B' + 1$. If $B > B_{\max}$, stop. Otherwise, calculate τ_j^* for a further $B' + 1$ bootstrap samples and set $B' = B$. Then return to step 2.

The rule in step 3 is essentially arbitrary, but it is very simple, and it ensures that $\alpha(B+1)$ is an integer if $\alpha(B'+1)$ is. It is easy to see how this procedure will work. When $p^*(\hat{\tau})$ is not close to α, it will usually terminate after one or two rounds with an estimate $\hat{p}^*(\hat{\tau})$ that is relatively inaccurate, but clearly different from α. When $p^*(\hat{\tau})$ is reasonably close to α, the procedure will usually terminate after several rounds with an estimate $\hat{p}^*(\hat{\tau})$ that is fairly accurate. When $p^*(\hat{\tau})$ is very close to α, it will usually terminate with $B = B_{\max}$ and a very accurate estimate $\hat{p}^*(\hat{\tau})$.

Occasionally, especially in this last case, the procedure will make a mistake, in the sense that $\hat{p}^*(\hat{\tau}) < \alpha$ when $p^*(\hat{\tau}) > \alpha$, or *vice versa*. In such cases, simulation randomness causes the result of a feasible bootstrap test at level α to be different from the (infeasible) result of the ideal bootstrap test. If $B_{\max} = \infty$, the probability of such conflicts between the feasible and ideal tests is bounded above by β. In practice, with $B_{\max}$ finite, the probability can be higher than β. However, the magnitude of the difference between $\hat{p}^*(\hat{\tau})$ and $p^*(\hat{\tau})$ in the case of a conflict is bound to be very small. In terms of the tradeoff between conflicts and computing time, it is desirable to keep β small in order to avoid conflicts when B is still small, but, for large B, the probability of conflicts can be reduced only by increasing $B_{\max}$, with a consequent increase in expected computing time.

The procedure just described could easily be modified to handle more than one value of α, if desired. For example, we might be interested in tests at

both the .01 and .05 levels. Then step 2 would be modified so that we would stop only if $\hat{p}^*(\hat{\tau}) > .05$ and we could reject the hypothesis that $p^*(\hat{\tau}) \leq .05$, or if $\hat{p}^*(\hat{\tau}) < .01$ and we could reject the hypothesis that $p^*(\hat{\tau}) \geq .01$, or if $.01 < \hat{p}^*(\hat{\tau}) < .05$ and we could reject both the hypothesis that $p^*(\hat{\tau}) \leq .01$ and the hypothesis that $p^*(\hat{\tau}) \geq .05$.

4. THE PERFORMANCE OF THE PRETEST PROCEDURE

In order to investigate how this procedure works in practice, we conducted several simulation experiments, with two million replications each, based on the model (3), with different values of γ, $B_{\min}$, $B_{\max}$, and β. The same sequence of random numbers was used to generate the data for all values of these parameters. $B_{\min}$ was normally 99, and α was always .05. Because it is extremely expensive to evaluate the binomial distribution directly when B is large, we used the normal approximation to the binomial whenever $\alpha B \geq 10$. Since (3) is so simple that the bootstrap distribution is known analytically, we can evaluate the ideal bootstrap P value $p^*(\hat{\tau})$ for each replication.

Table 1 shows results for four different values of γ. When $\gamma = 0$, so that the null hypothesis is true, B^*, the average number of bootstrap samples used by the procedure, is quite small. As expected, reducing β and increasing $B_{\max}$ both cause B^* to increase. As can be seen from the column headed "Conflicts", the procedure does indeed yield very few cases in which the feasible bootstrap test yields a different result from the ideal test. For all values of β considered, most conflicts occur when the procedure terminates with $B^* = B_{\max}$, which implies that $p^*(\hat{\tau})$ is very near α and is estimated very accurately. The last column, headed "Av. Diff.", shows the average absolute difference between $\hat{p}^*(\hat{\tau})$ and $p^*(\hat{\tau})$ in the few cases where there was a conflict. It is clear that, even when the procedure yields the "wrong" answer, the investigator is not likely to be seriously misled.

The average number of bootstrap samples is higher for $\gamma = 3$ than for $\gamma = 0$, higher again for $\gamma = 1$, and higher still for $\gamma = 2$. This reflects the way in which the proportion of the $p^*(\hat{\tau})$ near .05 depends on γ. When B^* is higher, there tend to be more cases in which the feasible and ideal bootstrap tests yield different results, because the procedure terminates more frequently with $B = B_{\max}$. It is clear that the procedure gives rise to some power loss relative to the ideal test, but it is always very small, less than .0007 in the worst case. All of the choices of β and $B_{\max}$ that were investigated appear to yield acceptable results, and it is difficult to choose among them. We tentatively recommend setting $B_{\max} = 12,799$ and choosing either $\beta = .01$ or $\beta = .001$. Using the smaller value of β modestly reduces the number of conflicts and substantially reduces the average size of the conflicts that do occur, but it also seems to reduce power slightly in two cases.

TABLE 1 Bootstrap Tests with B Chosen by Pretest

γ	$B_{\min}$	$B_{\max}$	β	B^*	Rej. B^∞	Rej. B^*	Conflicts	Av. Diff.
0.0	99	12,799	0.01	325.1	0.04982	0.04988	0.0017	0.0087
0.0	99	6,399	0.001	318.8		0.04986	0.0022	0.0037
0.0	99	12,799	0.001	420.9		0.04984	0.0015	0.0031
0.0	99	12,799	0.0001	491.0		0.04986	0.0015	0.0025
0.0	439	439		439.0		0.04998	0.0083	0.0132
1.0	99	12,799	0.01	1055.5	0.28861	0.28865	0.0073	0.0091
1.0	99	6,399	0.001	1033.5		0.28844	0.0093	0.0038
1.0	99	12,799	0.001	1472.5		0.28847	0.0065	0.0031
1.0	99	12,799	0.0001	1771.0		0.28840	0.0066	0.0025
1.0	1,499	1,499		1499.0		0.28790	0.0192	0.0070
2.0	99	12,799	0.01	1368.3	0.75479	0.75458	0.0094	0.0091
2.0	99	6,399	0.001	1404.3		0.75411	0.0118	0.0038
2.0	99	12,799	0.001	1973.9		0.75434	0.0085	0.0030
2.0	99	12,799	0.0001	2409.1		0.75450	0.0085	0.0025
2.0	1,999	1,999		1999.0		0.75279	0.0212	0.0061
3.0	99	12,799	0.01	566.6	0.96706	0.96683	0.0029	0.0087
3.0	99	6,399	0.001	705.1		0.96663	0.0036	0.0038
3.0	99	12,799	0.001	885.0		0.96684	0.0026	0.0030
3.0	99	12,799	0.0001	1120.1		0.96688	0.0026	0.0025
3.0	899	899		899.0		0.96449	0.0101	0.0091

B^* is the average value of B that was finally chosen.

"Rej. B^∞" is the proportion of replications for which the null hypothesis was rejected at the .05 level according to the t distribution.

"Rej. B^*" is the proportion of replications for which the null hypothesis was rejected at the .05 level according to the bootstrap test, based on whatever value of B was finally used.

"Conflicts" is the proportion of replications for which the ideal bootstrap test and the feasible bootstrap test yielded different inferences.

"Av. Diff." is the average absolute difference between the ideal and feasible bootstrap P values, for replications on which a conflict occurred.

The last line in the table for each value of γ shows what happens when we choose a fixed B slightly larger than the B^* observed for the recommended values of β and $B_{\max}$. There are far more conflicts when a fixed B is used, and they are on average much larger, because they are based on much less accurate estimates of $p^*(\hat{\tau})$. There is also substantially more power loss. Thus it appears that, holding expected computer time constant, the pretesting procedure works very much better than using a fixed value of B.

It is easy to understand why the pretesting procedure works well. When the null hypothesis is true, B can safely be small, because we are not concerned about power at all. Similarly, when the null is false and test power is extremely high, B does not need to be large, because power loss is not a serious issue. However, when the null is false and test power is moderately high, B needs to be large in order to avoid loss of power. The pretesting procedure tends to make B small when it can safely be small and large when it needs to be large.

5. AN ALTERNATIVE PROCEDURE

In a very recent paper, Andrews and Buchinsky (1998), hereafter referred to as A-B, propose another method for determining B. Their approach is based on the fact that, according to the normal approximation to the binomial distribution, if B bootstrap samples are used, then

$$B^{1/2}\big(\hat{p}^*(\hat{\tau}) - p^*(\hat{\tau})\big) \sim N\big(0, p(1-p)\big),$$

conditional on the randomness in the data. They wish to choose B so that the absolute value of the proportional error in $\hat{p}^*(\hat{\tau})$ exceeds some value d, that will generally be considerably less than 1, with probability ρ. If we write $p \equiv p^*(\hat{\tau})$ for simplicity, this implies that

$$B = \text{int}\left(\frac{\chi^2_{1-\rho}(1-p)}{pd^2}\right), \tag{4}$$

where $\chi^2_{1-\rho}$ is the $1-\rho$ quantile of the $\chi^2(1)$ distribution. (A-B use somewhat different notation, which would conflict with the notation we use in this paper.)

Because (4) depends on p, which is unknown, the A-B procedure involves three steps. In the first step, it uses the (known) asymptotic distribution of τ to compute an asymptotic P value for $\hat{\tau}$. This asymptotic P value is then used in (4) to calculate a preliminary number of bootstrap samples, B_1, and B_1 bootstrap samples are then drawn. The bootstrap P value computed from them is then used in (4) to calculate a final number of bootstrap samples, B_2. If $B_2 < B_1$, the procedure terminates with $B = B_1$. Otherwise, a further $B_2 - B_1$ bootstrap samples are drawn.

The above description of the A-B procedure suggests that all the user need choose in advance is d and ρ. However, a little experience suggests that it is also necessary to pick $B_{\max}$, since, as is clear from (4), B_1 will be extremely large when the asymptotic P value is near zero, as will B_2 when the first-stage bootstrap P value is near zero. Moreover, it is desirable to modify the procedure slightly so that B_1 and B_2 both satisfy the condition that $\alpha(B+1)$ is an integer when $\alpha = .05$, and this implies that $B \geq 19$.

Unlike the procedure we have proposed, the A-B procedure is intended to give a reasonably accurate estimate of p^* whether or not p^* is close to α. But achieving this goal entails a penalty in terms of our criteria of computing cost, power loss, and the number and magnitude of conflicts between the ideal and feasible bootstrap tests. To demonstrate these features of the A-B procedure, we performed a number of simulation experiments, comparable to those in Table 1. In these experiments, we tried several values of d and ρ and found that $d = .20$ and $\rho = .05$ seemed to provide reasonable results when $\gamma = 0$. This therefore became our baseline case. In all the experiments, we set $B_{\min} = 19$ and $B_{\max} = 12,799$.

TABLE 2 Bootstrap Tests with B Chosen by A-B Procedure

γ	d	ρ	δ	B_1	B_2	B^*	Rej. B^∞	Rej. B^*	Conflicts	Av. Diff.
0.0	.10	.05	1.0	1366.9	1370.3	1395.8	0.04982	0.04985	0.0019	0.0031
0.0	.20	.05	1.0	480.7	484.9	501.5		0.04986	0.0038	0.0060
0.0	.20	.10	1.0	365.2	370.0	384.2		0.04979	0.0044	0.0071
0.0	.20	.05	0.5	134.3	499.0	499.0		0.04963	0.0040	0.0064
0.0	.20	.05	2.0	1673.5	481.8	1673.6		0.04989	0.0016	0.0025
1.0	.10	.05	1.0	4982.9	4988.0	5043.5	0.28861	0.28845	0.0084	0.0031
1.0	.20	.05	1.0	2200.3	2212.9	2269.5		0.28810	0.0162	0.0060
1.0	.20	.10	1.0	1738.9	1754.1	1808.1		0.28774	0.0190	0.0071
1.0	.20	.05	0.5	601.0	2265.8	2265.8		0.28706	0.0171	0.0064
1.0	.20	.05	2.0	6032.2	2202.9	6032.2		0.28867	0.0067	0.0025
2.0	.10	.05	1.0	10222.5	10224.7	10279.0	0.75479	0.75424	0.0107	0.0031
2.0	.20	.05	1.0	6113.6	6132.1	6246.6		0.75291	0.0208	0.0060
2.0	.20	.10	1.0	5119.0	5145.6	5267.7		0.75191	0.0244	0.0070
2.0	.20	.05	0.5	1966.4	6216.9	6216.9		0.75124	0.0220	0.0064
2.0	.20	.05	2.0	11256.2	6116.5	11256.2		0.75470	0.0085	0.0025
3.0	.10	.05	1.0	12367.8	12367.1	12386.2	0.96706	0.96675	0.0033	0.0031
3.0	.20	.05	1.0	9680.2	9689.6	9804.4		0.96576	0.0064	0.0060
3.0	.20	.10	1.0	8624.9	8644.1	8788.2		0.96540	0.0076	0.0070
3.0	.20	.05	0.5	3969.0	9732.0	9732.0		0.96527	0.0071	0.0065
3.0	.20	.05	2.0	12659.4	9678.6	12659.4		0.96685	0.0027	0.0025

δ is a factor by which the test statistic is multiplied.

B_1 is the average number of bootstraps from step 1 of the A-B procedure.

B_2 is the average number of bootstraps from step 2 of the A-B procedure.

B^* is the average number of bootstraps in total.

See also the notes to Table 1.

The results of our simulations are presented in Table 2. Comparing these results with those in Table 1 shows that the A-B procedure performs

much less well than the pretesting procedure. Either it achieves similar performance based on far more bootstrap samples (for example, for $\gamma = 2$, compare A-B with $d = .10$ and $\rho = .05$ with any of the results in Table 1 except those with $\beta = .01$), or else it achieves much worse performance based on a similar or larger number of bootstrap samples (for example, for $\gamma = 1$, compare A-B with $d = .20$ and $\rho = .10$ with the other procedure with $B_{\max} = 12,799$ and $\beta = .0001$).

Most of our results actually show the A-B procedure in an unrealistically good light, because the asymptotic P value used to determine B_1 is correct. We therefore ran some experiments in which $\hat{\tau}$ was multiplied by a positive factor δ. When $\delta < 1$, the asymptotic test underrejects, and when $\delta > 1$, it overrejects. These errors cause B_1 to be chosen poorly. As can be seen from Table 2, overrejection causes the A-B procedure to use more bootstrap samples than it should, and underrejection causes it to lose power and have more conflicts, while only slightly reducing the average number of bootstrap samples. Note that multiplying $\hat{\tau}$ by any positive constant has absolutely no effect on the performance of a bootstrap test with B fixed or on a bootstrap test that uses our procedure to choose B.

6. FINAL REMARKS

An unavoidable feature of bootstrap testing is the need to choose the number of bootstrap samples, B. In Section 2, we discussed the loss of power that can occur when B is too small. In Section 3, we proposed a simple pretesting procedure designed to ensure that, for one or more chosen levels, feasible bootstrap tests (that is, ones with finite B) yield almost the same results as ideal bootstrap tests, while keeping B relatively small. We showed in Section 4 that this procedure works substantially better than using a fixed number of bootstrap samples. Finally, in Section 5, we showed that it also works much better than another procedure for choosing B that has recently been proposed.

ACKNOWLEDGEMENTS

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Joel Horowitz, Don Andrews, several referees, and numerous seminar participants for comments on earlier work.

REFERENCES

Andrews, D. W. K. and M. Buchinsky (1998). "On the number of bootstrap repetitions for bootstrap standard errors, confidence intervals, confidence regions, and tests," Cowles Foundation Discussion Paper No. 1141R, Yale University, revised.

Barnard, G. A. (1963). "Contribution to discussion," *J. R. Statist. Soc. B*, 25, 294.

Beran, R. (1988). "Prepivoting test statistics: a bootstrap view of asymptotic refinements," *J. Amer. Statist. Ass.*, 83, 687–697.

Davidson, R. and J. G. MacKinnon (1998). "Graphical methods for investigating the size and power of hypothesis tests," *Manch. School*, 66, 1–26.

Davidson, R. and J. G. MacKinnon (1999). "The size distortion of bootstrap tests," *Econometric Theory*, 15, forthcoming.

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

Dufour, J.-M. and J. F. Kiviet (1998). "Exact inference methods for first-order autoregressive distributed lag models," *Econometrica*, 66, 79–104.

Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses," *Ann. Math. Statist.*, 28, 181–187.

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 of accuracy and power of Monte Carlo tests," *J. R. Statist. Soc. B*, 51, 459–467.

Hope, A. C. A. (1968). "A simplified Monte Carlo significance test procedure," *J. R. Statist. Soc. B*, 30, 582–598.

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

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

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

Marriott, F. H. C. (1979). "Barnard's Monte Carlo tests: How many simulations?" *Appl. Statist.*, 28, 75–77.