Bootstrap Confidence Intervals Based on Inverting Hypothesis Tests

by

Russell Davidson

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

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

russell@ehess.cnrs-mrs.fr

Abstract

Most confidence intervals, whether based on asymptotic theory or the bootstrap, are implicitly based on inverting a Wald test. Since Wald test statistics are not invariant under nonlinear reparametrizations of the restrictions they test, confidence intervals based on them are not invariant either. This fact explains the well-known non invariance of bootstrap confidence intervals obtained by Hall's percentile-t method. Davidson and MacKinnon (1999) show that bootstrap inference can be improved if the bootstrapped test statistic is asymptotically independent of the bootstrap data-generating process. In this note, it is shown for a simple AR(1) model that greatly improved coverage accuracy of confidence intervals can be obtained by explicitly inverting a set of bootstrap hypothesis tests for each of which the bootstrap data-generating process is asymptotically independent of the bootstrapped statistic.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. This note is based on a comment on a paper, *Recent Developments on Bootstrapping Time Series*, by Berkowitz and Kilian. The paper and the comment were published in 2000 in *Econometric Reviews*.

March, 2000

It bears witness to the enormous interest in bootstrap methods in econometrics, and the rate of progress in their application, that Berkowitz' and Kilian's paper should appear such a short time after the paper by Li and Maddala (1996) on bootstrapping time series models in *Econometric Reviews*, with new and interesting material not covered in the previous, very widely cited, and influential piece.

In this comment, I would like to explore what at first sight seems to be a point of disagreement between Li and Maddala on the one hand, and Berkowitz and Kilian on the other, namely the importance of using pivotal quantities for effective bootstrapping. My own position has always been that of Li and Maddala: One wastes the potential of the bootstrap for accurate inference if one bootstraps quantities, like parameter estimates, that are not asymptotically pivotal. However, Berkowitz and Kilian present some pretty convincing evidence that this position is not universally justifiable.

I suppose that no one would dispute that bootstrap inference is better when the quantity bootstrapped is closer to being pivotal. But Berkowitz and Kilian argue that, if one wishes to perform inference on things like impulse response coefficients, it seems to be the case that, for realistic sample sizes, the estimates of the variances of these coefficients are so noisy that studentizing leads to quantities that are farther from being pivotal than the estimated response coefficients themselves. In that case, it follows that we will do better by bootstrapping the raw response coefficients rather than studentized versions of them. Of course, this is not an argument against bootstrapping pivotal or nearly pivotal quantities. It is an argument against studentizing when the result of studentizing is no closer to pivotal than the quantity studentized.

I think it is profitable in the analysis of this phenomenon to think of confidence intervals as based on a set of hypothesis tests. A value belongs to a confidence interval of nominal coverage $1-\alpha$ if a test of the hypothesis that this value is the true one is not rejected by a test of nominal significance level α. Thus a confidence interval can be thought of as the result of "inverting" the set of hypothesis tests. Conversely, any rule for constructing confidence intervals implicitly defines a set of tests. Formally, if we wish to test the hypothesis that a scalar parameter θ equals some specific value θ_0, the P value based on a set of confidence intervals is one minus the nominal coverage of the confidence interval for which θ_0 is a boundary point.

Although the duality described above between confidence intervals and hypothesis tests is perfectly general in theory, in practice most confidence intervals are obtained by inverting Wald tests. The most common sort of confidence interval is obtained by computing a parameter estimate $\hat{\theta}$ and its corresponding standard error $\sigma(\hat{\theta})$, and then forming an interval of the form $[\hat{\theta} - c\sigma(\hat{\theta}), \hat{\theta} + c\sigma(\hat{\theta})]$, where c is a critical value for the assumed distribution of the studentized statistic $W \equiv (\hat{\theta} - \theta)/\sigma(\hat{\theta})$. Often the assumed distribution is just standard normal, or perhaps the Student's t distribution when there are few degrees of freedom. If bootstrap percentile-t confidence intervals are used, the distribution of which c is a critical value is obtained by bootstrapping W, and, commonly, different values of c are used for the lower and upper limits.

If the above confidence interval has nominal coverage $1 - \alpha$, then the hypothesis that the true parameter is equal to any given θ will be rejected at nominal level α if and only if $\theta \notin [\hat{\theta} - c\sigma(\hat{\theta}), \hat{\theta} + c\sigma(\hat{\theta})]$, that is,

$$\left|(\hat{\theta} - \theta)/\sigma(\hat{\theta})\right| = |W| > c.$$

This rejection rule is thus based on the Wald statistic W, which is supposed to follow a standard normal or t distribution under the null for asymptotic tests, or a distribution found by simulation for a bootstrap test.

It is well known – see Gregory and Veall (1985) and Phillips and Park (1988) – that Wald tests are not invariant under nonlinear reparametrizations. In fact, as shown by Lafontaine and White (1986), any prespecified value at all can be obtained as the Wald statistic for a given hypothesis with a suitable nonlinear reparametrization. It follows from this that bootstrap percentile-t confidence intervals, for which the Wald statistic W is bootstrapped, are not invariant under nonlinear reparametrizations any more than the underlying statistic. The non-invariance of percentile-t intervals is well known, but it does not seem to be widely appreciated that it is just a consequence of the non-invariance of the Wald test.

In the sort of dynamic models considered by Berkowitz and Kilian, the impulse responses are just nonlinear functions of the parameters of the underlying ARMA process that generates the data. Confidence intervals based on estimates of the impulse responses, with or without studentizing, are therefore based on nonlinear reparametrizations of the ARMA process, and can, as is seen very clearly in the simulation experiments of Berkowitz and Kilian, lead to very erratic behavior. The moral of this tale is presumably that it would be better to construct confidence intervals for the impulse responses by first constructing confidence regions for the ARMA parameters, and then projecting these on to the impulse responses by means of the appropriate nonlinear transformation.

Beran (1988) showed that bootstrap inference is refined when the quantity bootstrapped is nearly pivotal. Davidson and MacKinnon (1999) show that a further refinement, known to exist in a variety of seemingly unrelated circumstances, is more generally available if the quantity bootstrapped is nearly independent of the random quantities involved in setting up the bootstrap data-generating process (DGP). Such near independence is often easy to obtain by basing the bootstrap DGP on estimates obtained under the null hypothesis. Berkowitz and Kilian have shown here that studentizing is not necessarily enough to make quantities that are bootstrapped close to pivotal, although I suspect that this is a much less severe problem if only the underlying ARMA parameters, or studentized versions of them, are bootstrapped. Be that as it may, it is interesting to see if inference on ARMA parameters can be improved over that given by percentile-t intervals by using bootstrap DGPs which, being based on estimates under the null, are nearly independent of the parameter estimates.

An obstacle to this is that a confidence interval for a parameter θ is obtained by inverting a whole set of tests, since each possible value of θ corresponds to a different null hypothesis, with its own value for the test statistic. In constructing a percentile-t interval, this issue is finessed by bootstrapping only under the DGP

characterized by the estimate $\hat{\theta}$. Since $W = (\hat{\theta} - \theta)/\sigma(\hat{\theta})$ is in favorable circumstances nearly pivotal, its distribution is not much affected by the particular value of θ used for simulations. Clearly $\hat{\theta}$ itself is the most convenient value. But the bootstrapped statistic W and the bootstrap DGP based on $\hat{\theta}$ are in no way nearly independent, and so we cannot expect to benefit from the second refinement of Davidson and MacKinnon when we construct percentile-t intervals.

It is quite possible to construct a bootstrap confidence interval for which the simulations are performed using bootstrap DGPs that are nearly independent of the test statistic being bootstrapped. The procedure is more computationally intensive than for a percentile-t interval, but much less so than for a procedure using a double layer of bootstrapping. I illustrate the procedure for the very simple problem of finding a confidence interval for the autoregressive parameter ρ in the model

$$y_t = \rho y_{t-1} + u_t, \tag{1}$$

where, for simplicity, the error terms u_t are assumed to be independently distributed as $N(0, 1)$. Since the model is scale invariant, there is no loss of generality in setting the error variance to 1. The first step is to estimate (1) by OLS, thus obtaining $\hat{\rho}$ and its standard error $\sigma(\hat{\rho})$. For nominal coverage of 95%, we set the critical value c equal to 1.96, and thus obtain the limits of the asymptotic confidence interval, $\rho_\pm \equiv \hat{\rho} \pm 1.96\sigma(\hat{\rho})$. Next, bootstrap P values are computed for each of ρ_+ and ρ_- as follows. For the null that $\rho = \rho_+$, the test statistic is $\tau_+ \equiv (\hat{\rho} - \rho_+)/\sigma(\hat{\rho})$, and the bootstrap DGP that satisfies this null and is independent of τ_+ is $y_t^* = \rho_+ y_{t-1}^* + u_t^*$, where the u_t^* can be generated directly from $N(0, 1)$ or by resampling from the OLS residuals from running (1). For each bootstrap replication $j = 1, \ldots, B$, we run (1) using the simulated data y_t^* so as to obtain $\hat{\rho}_j^*$ and $\sigma(\hat{\rho}_j^*)$. The bootstrap statistic τ_j^* is computed as $(\hat{\rho}_j^* - \rho_+)/\sigma(\hat{\rho}_j^*)$. Finally, the bootstrap P value is

$$P_+ = \frac{1}{B} \sum_{j=1}^{B} I\big(|\tau_j^*| > |\tau_+|\big),$$

where $I(\cdot)$ is an indicator function, so that P_+ is the proportion of bootstrap replications for which the bootstrap statistic τ_j^* is more extreme than the statistic τ_+ computed with the original data. P_- is computed in just the same way with ρ_- and $\tau_- = (\hat{\rho} - \rho_-)/\sigma(\hat{\rho})$ in place of ρ_+ and τ_+ respectively.

A suitable root-finding algorithm can now be invoked to adjust the values of ρ_- and ρ_+ until the bootstrap P values are exactly equal to 0.05. I used a version of the `RTSAFE` algorithm of Press *et al* (1992), for which I simply assumed that the derivative of the P value with respect to $\rho_\pm$ was the standard normal density evaluated at 1.96. Despite the crudeness of this approximation, the algorithm converged reliably. A practical point here is that the *same* random numbers should be used for the bootstrap DGPs with different values of $\rho_\pm$. Otherwise, bootstrap randomness is enough to rule out convergence with reasonable values of B. The final bootstrap confidence interval is $[\rho_-, \rho_+]$ after convergence.

In order to test this procedure, I performed 200,000 replications with $\rho = 0.95$ and a sample size of 10. On each replication, I computed the asymptotic confidence

interval, Hall's bootstrap percentile and percentile-t intervals, and the interval described above, using $B = 399$ bootstrap replications. Although 399 is rather a small number, the averaging over 200,000 replications makes it entirely adequate for present purposes. It was necessary to choose a value of ρ close to unity and a very small sample size in order to produce significant coverage errors even for the asymptotic confidence interval. The results were as follows. The asymptotic interval covered the true value of 0.95 in 182,856 replications, the percentile interval in 186,455, the percentile-t in 188,802, and the new interval in 189,866 replications. If coverage were equal to nominal, the true value would be covered 190,000 times. The standard error associated with these numbers is roughly the square root of $0.05 \times 0.95 \times 200,000$, or nearly 100. Thus the coverage error of the new interval is not significant at conventional levels even with 200,000 replications.

It is probably unrealistic to expect such reliability with more complicated ARMA models, but it seems reasonable to conclude that this technique yields better inference than percentile or percentile-t intervals on account of the near independence of the statistic and the bootstrap DGP. In terms of computing cost, on average 14 iterations were needed for ρ_- and ρ_+, 6.5 for ρ_-, and 7.5 for ρ_+. Computing time is not as much as 15 times that for the percentile-t interval, since the random numbers for the bootstrap are computed only once, and there are other small economies of scale in the computation. This is a very reasonable price to pay for such greatly improved inference. It will be interesting to see if the improvement is as great in models more complicated than the toy model considered here.

References

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 (1999). "The size distortion of bootstrap tests," *Econometric Theory*, 15, forthcoming.

Gregory, A. W., and M. R. Veall (1985). "On formulating Wald tests for nonlinear restrictions," *Econometrica*, **53**, 1465–68.

Lafontaine, F., and K. J. White (1986). "Obtaining any Wald statistic you want," *Economics Letters*, **21**, 35–40.

Li, H., and G. S. Maddala (1996). "Bootstrapping Time Series Models," *Econometric Reviews*, **15**, 297–318.

Phillips, P. C. B., and J. Y. Park (1988). "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, **56**, 1065–83.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, (1992). *Numerical Recipes in FORTRAN*, Second edition, Cambridge, Cambridge University Press.