Bootstrap Methods in Econometrics

James G. MacKinnon

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

jgm@econ.queensu.ca
http://www.econ.queensu.ca/faculty/mackinnon/

September, 2005

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada.

Introduction:

Bootstrap methods involve estimating a model many times using simulated data. Then quantities computed from the simulated data are used to make inferences from the actual data.

Why have bootstrap methods become popular?

- Drop in cost of numerical computation. 200 times less than a decade ago; perhaps 100,000 times less than two decades ago.
- **Bootstrapping** is widely thought to be easy.
- Bootstrapping is widely believed to work well.

There are many different **bootstrap methods**.

- Some bootstrap methods are easy.
- Some bootstrap methods work extraordinarily well in some cases.
- Bootstrap methods do not always work well.
- Choosing among different bootstrap methods is often not easy.

Hypothesis Testing:

$\hat{\tau}$: realized value of a test statistic τ.

$F(\tau)$: **cumulative distribution function**, or **CDF**, of τ.

If we reject the null hypothesis whenever $\hat{\tau}$ is abnormally large, and we know $F(\tau)$, we can either:

1. Calculate a **critical value** at level α, say c_α, defined by

$$1 - F(c_\alpha) = \alpha, \tag{1}$$

   and reject the null whenever $\hat{\tau} > c_\alpha$. For example, if $\alpha = .05$, and $F(\cdot)$ is the $\chi^2(1)$ distribution, $c_\alpha = 3.84$.

2. Calculate the ***P* value**

$$P(\hat{\tau}) = 1 - F(\hat{\tau}), \tag{2}$$

   and reject whenever $P(\hat{\tau}) < \alpha$.

These two procedures yield identical inferences.

When do not know $F(\tau)$, usual approach is to replace it by the **asymptotic CDF** $F^\infty(\tau)$.

Bootstrap Testing:

The bootstrap provides another way to approximate $F(\tau)$, which may provide a better approximation.

Bootstrap can be used even when τ is complicated to compute and/or does not have a known asymptotic distribution.

For example, τ might be the maximum of several (dependent) test statistics, or the minimum of several (asymptotic) P values.

- We might be performing a number of specification tests and wish to control the size of the entire procedure.
- A well-known example is testing for structural change with an unknown break point (Hansen, 2000).
- The test statistic may follow a non-standard distribution asymptotically, as many test statistics for unit roots and cointegration do.

For the bootstrap to work well (in theory), the test statistic must have an asymptotic distribution; we do not need to know that distribution.

In order to perform a bootstrap test, we must generate B **bootstrap samples**, indexed by j, that satisfy the null hypothesis.

For each bootstrap sample, compute a **bootstrap test statistic** τ_j^*, usually by the same procedure used to calculate $\hat{\tau}$ from the real sample.

The **bootstrap *P* value** is

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

where $I(\cdot)$ is the **indicator function**. This can also be written as

$$p^*(\hat{\tau}) = 1 - \hat{F}^*(\hat{\tau}), \tag{4}$$

where $\hat{F}^*(\tau)$ is the **empirical distribution function**, or **EDF**, of the τ_j^*.

Bootstrap P value (4) looks just like true P value (2), but with $\hat{F}^*(\hat{\tau})$ replacing $F(\hat{\tau})$.

Monte Carlo Tests:

In an important special case, bootstrap tests are **exact**. That is, the probability of rejecting the null at level α is precisely α.

For this result to hold, we need two conditions:

1. The test statistic τ is **pivotal**, which means that its distribution does not depend on any unknown parameters.
2. The number of bootstrap samples B is such that $\alpha(B+1)$ is an integer.

Proof:

- By 1, $\hat{\tau}$ and the τ_j^* all come from the same distribution.
- Imagine sorting all $B+1$ test statistics. If $\hat{\tau}$ is one of the largest $\alpha(B+1)$ statistics, we reject the null.
- Under the null, because of 2, this happens with probability α.
- For example, if $B = 999$ and $\alpha = .05$, we reject whenever $\hat{\tau}$ is one of the largest 50 test statistics.

Applications of Monte Carlo tests include many specification tests in linear regression models with fixed regressors and normal errors:

- Durbin-Watson tests and other tests for serial correlation
- Tests for ARCH errors and other forms of heteroskedasticity
- Jarque-Bera tests and other tests for skewness and kurtosis

These test statistics just depend on the OLS residuals $\boldsymbol{M_X}\boldsymbol{\varepsilon}$, where $\sigma\boldsymbol{\varepsilon} = \boldsymbol{u}$, on $\boldsymbol{M_X} = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, and perhaps on $\boldsymbol{X}$ directly.

If we know $\boldsymbol{X}$ and distribution of the ε_t, we can generate bootstrap test statistics that follow the same distribution as the actual test statistic under the null.

Under normality, draw the ε_t as independent standard normal variates.

When normality assumption is false, Monte Carlo tests for serial correlation should still be very accurate (though not exact), but Monte Carlo tests for heteroskedasticity might be quite inaccurate.

In some cases, it might make sense to assume some other distribution.

- It is possible to perform exact Monte Carlo tests even when $\alpha(B+1)$ is not an integer — see Racine and MacKinnon (2004) — but it is only worth the trouble if simulation is very expensive.
- Penalty for using a small number of bootstrap samples is loss of power, not loss of exactness.
- Even $B = 19$ is valid for a test at the .05 level, and $B = 99$ for a test at the .01 level.
- Unless computation is expensive, $B = 999$ is often a good choice.

Other references on Monte Carlo tests include:

Dwass (1957);
Jöckel (1986);
Dufour and Khalaf (2001);
Dufour, Khalaf, Bernard, and Genest (2004).

Properties of Bootstrap Tests:

When a test statistic is not pivotal, bootstrap tests are not exact, because $F^*(\tau)$ differs from $F(\tau)$.

- This problem goes away as $n \to \infty$ if τ is **asymptotically pivotal**, but it does not go away as $B \to \infty$.
- When $F(\tau)$ is not very sensitive to unknown parameters or distributions, bootstrap tests should work well.
- When $F(\tau)$ is sensitive to values of unknown parameters, or distributions, and they are estimated inefficiently and/or with large bias, bootstrap tests may work very badly.
- When there are alternative test statistics (e.g. LR, LM, and W), we should use the one that is closest to being pivotal.
- Different bootstrap methods applied to the same test statistic may have very different finite-sample properties.

- If τ is a test statistic and $\tau' = g(\tau)$ for any monotonic function $g(\cdot)$, then a bootstrap test based on τ' will yield exactly the same inferences as a bootstrap test based on τ. (e.g. F and LR tests in regression models)
- Bootstrap tests are *not* less powerful than asymptotic tests, on a properly size-adjusted basis, provided B is reasonably large.
- Comparing power of tests that are not exact is *extremely* difficult.

See Horowitz and Savin (2000), Davidson and MacKinnon (1999, 2005a) and MacKinnon (2002).

Bootstrapping Regression Models:

Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad \mathrm{E}(u_t \mid \boldsymbol{X}_t) = 0, \quad u_t \sim \mathrm{IID}(0, \sigma^2), \tag{5}$$

where there are n observations and k regressors. Regressors may include lagged dependent variables, but y_t is not explosive and does not have a unit root.

There are *many* ways to specify **bootstrap data generating processes**, or **bootstrap DGPs**, for this model. Some involve quite strong assumptions, others very weak ones.

In general, making stronger assumptions results in better performance if those assumptions are satisfied, but it leads to asymptotically invalid inferences if they are not.

- Are the errors independent?
- Are the errors identically distributed?

If the answer to both questions is yes, we can generally make very accurate inferences using a particular form of bootstrapping.

1. Residual bootstrap

Requires that the errors be independent of contemporaneous regressors and IID, but with minimal distributional assumptions.

Use OLS to obtain $\hat{\boldsymbol{\beta}}$ and the $\hat{u}_t$. Optionally, rescale residuals so that they have the correct variance. Simplest rescaled residual is

$$\ddot{u}_t \equiv \left(\frac{n}{n-k}\right)^{1/2} \hat{u}_t. \tag{6}$$

Another method uses the diagonals of the **hat matrix** $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. It may work a bit better when some observations have high leverage; see below. Details are given in Davidson and MacKinnon (2005b).

Generate a typical observation of the bootstrap sample as

$$y_t^* = \boldsymbol{X}_t\hat{\boldsymbol{\beta}} + u_t^*, \quad u_t^* \sim \text{EDF}(\ddot{u}_t). \tag{7}$$

The u_t^* are drawn from the **empirical distribution function**, or **EDF**, which assigns probability $1/n$ to each of the $\ddot{u}_t$.

The u_t^* are often said to be **resampled** from the $\ddot{u}_t$.

2. Parametric bootstrap

Assume that the error terms follow the normal distribution (or possibly some other distribution). Generate a typical observation of the bootstrap sample as

$$y_t^* = \boldsymbol{X}_t\hat{\boldsymbol{\beta}} + u_t^*, \quad u_t^* \sim \text{NID}(0, s^2). \tag{8}$$

Similar methods can be used with any model estimated by maximum likelihood, but their validity depends on the strong assumptions inherent in MLE.

- For inferences about regression coefficients, it generally makes very little difference whether we use residual bootstrap or parametric (normal) bootstrap.
- For inferences about other aspects of a model, such as possible, heteroskedasticity, it can make a large difference.

Restricted versus unrestricted estimates

Both these methods can easily be modified to impose restrictions on $\boldsymbol{\beta}$:

Just estimate the model under the null and use the restricted estimates $\tilde{\boldsymbol{\beta}}$ instead of the unrestricted estimates $\hat{\boldsymbol{\beta}}$.

- It is very important to use restricted parameter estimates in the bootstrap DGP when testing restrictions on $\boldsymbol{\beta}$. Otherwise, we must change the bootstrap test statistic; see below.
- Imposing the restrictions yields more efficient estimates, which makes bootstrap tests more accurate, because the bootstrap DGP is less random. See Davidson and MacKinnon (1999).
- It is OK to use unrestricted residuals instead of restricted residuals, even when parameters of bootstrap DGP are restricted.

3. Wild bootstrap

Specifically designed to handle heteroskedasticity in regression models. Proposed originally by Wu (1986).

The wild bootstrap DGP is

$$y_t^* = \boldsymbol{X}_t \hat{\boldsymbol{\beta}} + f(\hat{u}_t) v_t^*, \tag{9}$$

where $f(\hat{u}_t)$ is a transformation of the t^{th} residual $\hat{u}_t$, and v_t^* is a random variable with mean 0 and variance 1.

A good choice for $f(\cdot)$ is

$$f(\hat{u}_t) = \frac{\hat{u}_t}{(1 - h_t)^{1/2}}, \tag{10}$$

where h_t is the t^{th} diagonal of the hat matrix. These $f(\hat{u}_t)$ would have constant variance if the error terms were homoskedastic.

There are various ways to specify the distribution of the v_t^*. The simplest (but not the most popular) is

$$v_t^* = 1 \text{ with probability } \tfrac{1}{2}; \quad v_t^* = -1 \text{ with probability } \tfrac{1}{2}. \tag{11}$$

Thus each bootstrap error term can take on only two possible values!

Davidson and Flachaire (2001) have shown that wild bootstrap tests based on (11) usually perform better than other sorts of wild bootstrap test when the conditional distribution of the error terms is approximately symmetric.

The null hypothesis can (and should) be imposed if we are testing a hypothesis about $\boldsymbol{\beta}$.

Although it might seem that the wild bootstrap works only with cross-section data or static models, variants of it can be used with dynamic models. See Gonçalves and Kilian (2004).

4. Pairs bootstrap

Resample from the matrix with typical row $[y_t \;\; \boldsymbol{X}_t]$. We no longer condition on the $\boldsymbol{X}_t$, since each bootstrap sample now has a different $\boldsymbol{X}$ matrix. A typical observation of the bootstrap sample is $[y_t^* \;\; \boldsymbol{X}_t^*]$.

Proposed by Freedman (1981, 1984); see also Freedman and Peters (1984).

- The pairs (or **cases**) bootstrap is valid even when the errors display heteroskedasticity of unknown form.
- It works even for dynamic models. If regressors include lagged dependent variables, we treat them like any other element of $\boldsymbol{X}_t$.
- Pairs bootstrap can be applied to an enormous range of models.
- In the case of multivariate models, we can combine the pairs and residual bootstraps. Organize residuals as a matrix and apply the pairs bootstrap to its rows. This preserves cross-equation correlations.

Unfortunately, the pairs bootstrap has two major deficiencies:

- If the null hypothesis imposes restrictions on $\boldsymbol{\beta}$, the bootstrap DGP does not impose them. We must therefore modify the bootstrap test statistic so that it is testing something which is true in the bootstrap DGP.

$$\text{Actual test statistic:} \quad \frac{\hat{\beta}_1 - \beta_1^0}{\text{s.e.}(\hat{\beta}_1)} \tag{12}$$

$$\text{Bootstrap test statistic:} \quad \frac{\hat{\beta}_1^* - \hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1^*)} \tag{13}$$

- Compared to residual bootstrap (when it is valid) and wild bootstrap, pairs bootstrap does not yield very accurate results. There are both theoretical and simulation results on this point.

A Comparison of Several Methods:

We wish to test the null hypothesis that $\beta_2 = 0.9$ (not 0 or 1!) in the model

$$y_t = \beta_1 + \beta_2 y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \tag{14}$$

using the usual t statistic $(\hat{\beta}_2 - 0.9)/\text{se}(\hat{\beta}_2)$.

Five methods of inference:

- Asymptotic based on Student's t distribution.
- Residual bootstrap using restricted estimates and restricted rescaled residuals (RR).
- Residual bootstrap using unrestricted estimates and unrestricted rescaled residuals (UR).
- Pairs bootstrap.
- Wild bootstrap (Davidson-Flachaire method; simplest rescaling).

n: 10, 14, 20, 28, 40, 56, 80, 113, 160, 226, 320, 452, 640, 905, 1280.

100,000 replications, $B = 399$.

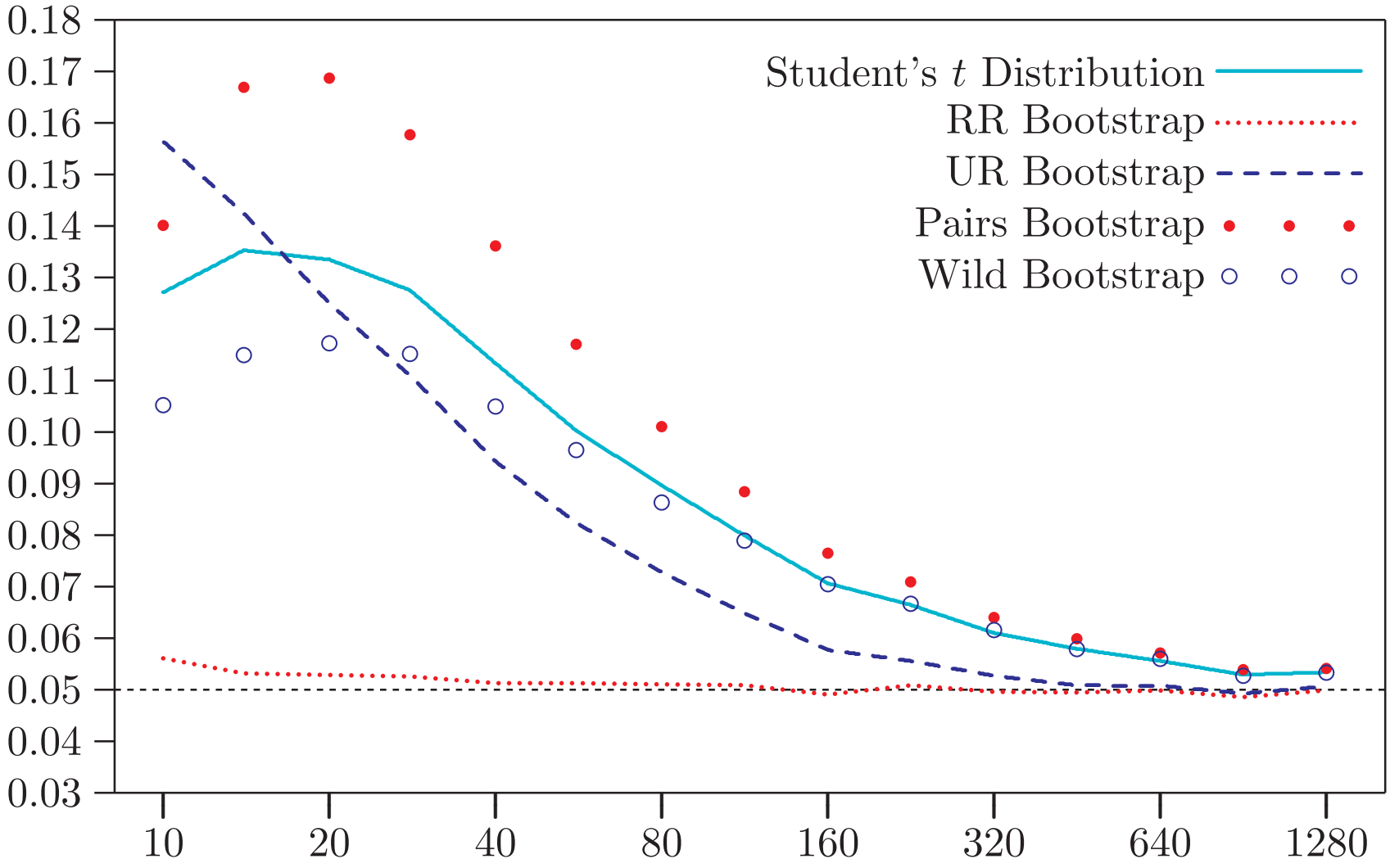


Figure 1. Rejection frequencies at .05 level under the null: IID errors

This example is unfair to the wild and pairs bootstraps, because the errors are IID. Suppose instead they follow the GARCH(1, 1) process

$$\sigma_t^2 \equiv \mathrm{E}(u_t^2) = \alpha_0 + \alpha_1 u_{t-1}^2 + \delta_1 \sigma_{t-1}^2, \tag{15}$$

with

$$\alpha_0 = 0.1, \quad \alpha_1 = 0.1, \quad \delta_1 = 0.8. \tag{16}$$

Instead of using ordinary t statistics, we now use the statistic

$$\frac{\hat{\beta}_2 - 0.9}{\mathrm{se_h}(\hat{\beta}_2)}, \tag{17}$$

where $\mathrm{se_h}(\hat{\beta}_2)$ is a heteroskedasticity-consistent standard error based on the HC_2 HCCME; see Davidson and MacKinnon (2004, Chapter 5).

Comparing Figures 1 and 2 makes it very clear that the "correct" bootstrap DGP to use depends on how the data are actually generated.

- Residual bootstraps (RR and UR) both work badly.
- Wild bootstrap works extraordinarily well.
- Pairs bootstrap also works well, but this may be misleading.

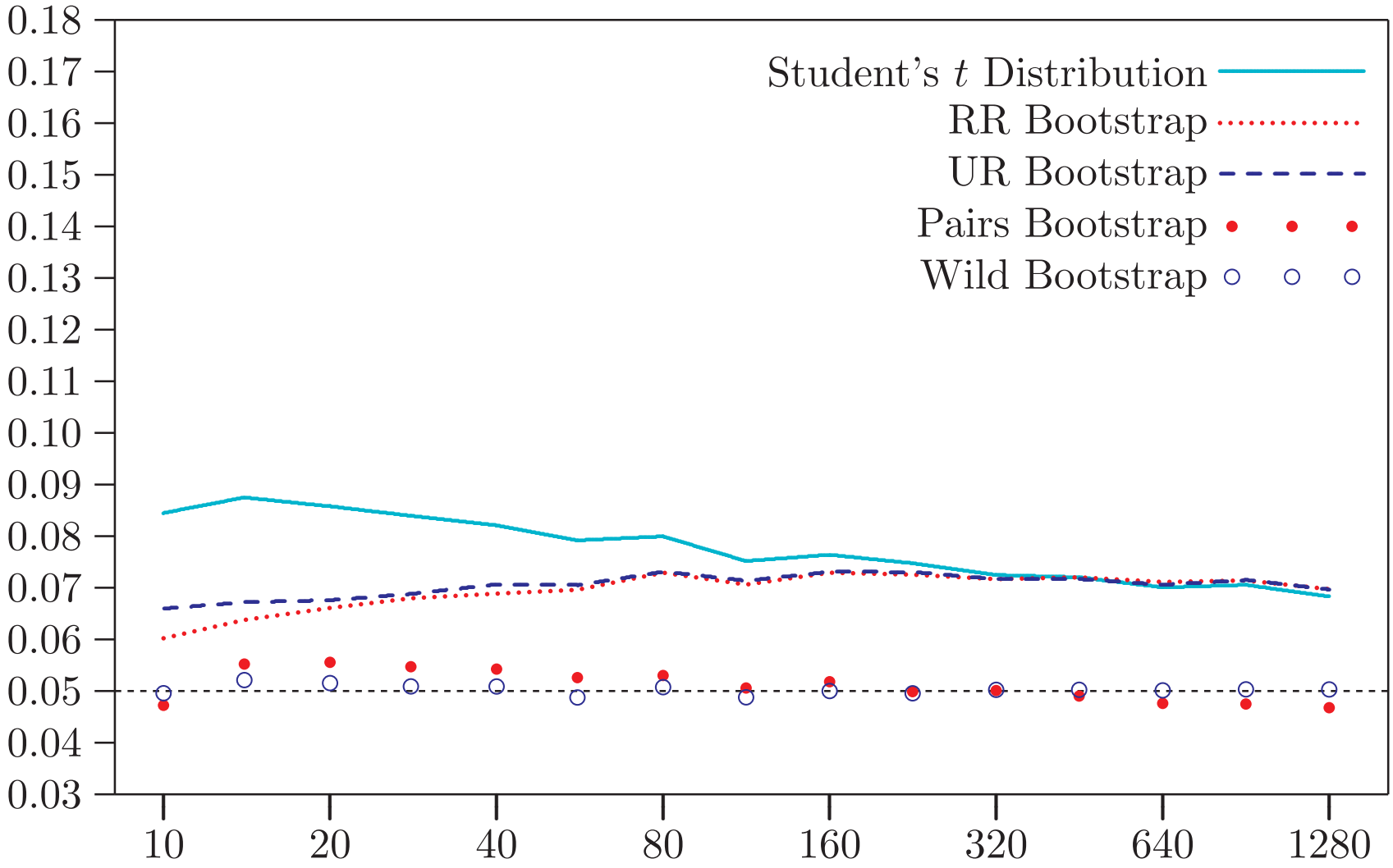


Figure 2. Rejection frequencies at .05 level under the null: GARCH errors

Bootstrap Standard Errors:

Original motivation for the bootstrap was to compute standard errors. See Efron (1979, 1982) .

This makes sense when other methods are conceptually or computationally difficult, are unreliable, or are not available at all.

If $\hat{\theta}$ is a parameter estimate, $\hat{\theta}_j^*$ is the estimate for the j^{th} bootstrap replication, and $\bar{\theta}^*$ is the mean of the $\hat{\theta}_j^*$, then the bootstrap standard error is

$$\text{se}^*(\hat{\theta}) = \left(\frac{1}{B-1} \sum_{j=1}^{B} (\hat{\theta}_j^* - \bar{\theta}^*)^2 \right)^{1/2}. \tag{18}$$

We can use $\text{se}^*(\hat{\theta})$ in the same way as we would use any other asymptotically valid standard error.

Warning! In the context of least squares, it makes no sense to bootstrap standard errors.

There are two standard covariance matrix estimators when the errors are independent:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{19}$$

$$\widehat{\mathrm{Var}}_{\mathrm{h}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{20}$$

Whatever the bootstrap DGP, the bootstrap covariance matrix is

$$\widehat{\mathrm{Var}}^*(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1} \sum_{j=1}^{B} (\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)^\top. \tag{21}$$

For a semiparametric bootstrap DGP that resamples residuals, this tends to

$$\sigma_*^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{22}$$

where σ_*^2 is the average variance of the bootstrap errors. This is valid if the errors are homoskedastic, but not otherwise.

For the pairs and wild bootstraps, (21) tends to something that has the same limit as the HCCME (20). The main difference is that it involves averages like

$$\frac{1}{B}\sum_{j=1}^{B}(\boldsymbol{X}_j^*)^\top \boldsymbol{X}_j^* \tag{23}$$

instead of the matrix $\boldsymbol{X}^\top\boldsymbol{X}$, and similarly for the expression in the middle of the sandwich. But these averages are very similar to $\boldsymbol{X}^\top\boldsymbol{X}$ for large B.

Bootstrap Confidence Intervals:

There are many ways to construct bootstrap confidence intervals for any given bootstrap DGP. See the extensive statistics literature; Davison and Hinkley (1997) is a good place to start.

1. In theory, the **percentile *t*** or **bootstrap *t*** method should work better than many other methods, in terms of the rate at which performance improves as n increases. It is said to offer **higher-order accuracy**; Hall (1992).

A percentile t confidence interval for θ at level $1 - \alpha$ is

$$\left[\hat{\theta} - \hat{s}_\theta \, t^*_{1-\alpha/2}, \;\; \hat{\theta} - \hat{s}_\theta \, t^*_{\alpha/2}\right], \tag{24}$$

where $\hat{s}_\theta$ is the standard error of $\hat{\theta}$, and t^*_δ is the δ quantile of the bootstrap t statistics

$$t^*_j = \frac{\hat{\theta}^*_j - \hat{\theta}}{\mathrm{se}(\hat{\theta}^*_j)}. \tag{25}$$

For example, if $\alpha = .05$ and $B = 999$, $t^*_{1-\alpha/2}$ will be number 975, and $t^*_{\alpha/2}$ will be number 25 in the sorted list of the t^*_j.

- This is like an ordinary confidence interval based on inverting a t statistic, but we use quantiles of the bootstrap distribution of the t^*_j instead of quantiles of the Student's t distribution.
- Percentile t method cannot be used if $\hat{s}_\theta$ cannot be calculated. It should not be used if $\hat{s}_\theta$ is unreliable, especially if strongly dependent on $\hat{\theta}$.

2. We can always calculate the bootstrap standard error (18) and construct a confidence interval based on the normal or Student's t distribution:

$$\left[\hat{\theta} - \mathrm{se}^*(\hat{\theta})\, t_{1-\alpha/2},\ \ \hat{\theta} + \mathrm{se}^*(\hat{\theta})\, t_{1-\alpha/2}\right]. \tag{26}$$

   In theory, this **simple bootstrap interval** works no better than asymptotic confidence interval and less well than percentile t.

   But it can be used when no standard error is available, B does not have to be large, and it often works quite well.

- A modified version performs **bias correction**. Replace $\hat{\theta}$ in (26) by

$$\check{\theta} = \hat{\theta} - (\bar{\theta}^* - \hat{\theta}) = 2\hat{\theta} - \bar{\theta}^*. \tag{27}$$

  Bias correction is often not a good idea, because it adds noise, but it can be helpful if the bias is severe and does not depend much on θ; see MacKinnon and Smith (1998).

- The percentile t method implicitly performs a sort of bias correction.

3. For both approaches, nonlinear transformations can be useful. Let $g(\cdot)$ be a monotonically increasing transformation. Thus

$$\phi = g(\theta) \quad \text{or, equivalently,} \quad \theta = g^{-1}(\phi). \tag{28}$$

Then the symmetric, simple bootstrap interval for ϕ,

$$\left[\hat{\phi} - \text{se}^*(\hat{\phi})\, t_{1-\alpha/2},\ \ \hat{\phi} + \text{se}^*(\hat{\phi})\, t_{1-\alpha/2}\right], \tag{29}$$

implies the nonsymmetric, transformed bootstrap interval for θ,'

$$\left[g^{-1}\left(\hat{\phi} - \text{se}^*(\hat{\phi})\, t_{1-\alpha/2}\right),\ \ g^{-1}\left(\hat{\phi} + \text{se}^*(\hat{\phi})\, t_{1-\alpha/2}\right)\right], \tag{30}$$

- The trick is to find a transformation such that $\hat{\phi}$ has an approximately symmetric distribution with a standard error that can estimated accurately using the bootstrap. If so, (29) should be a reliable interval, and hence also (30).
- Of course, using transformations may also help percentile t intervals, other bootstrap intervals, and asymptotic intervals.

A Disturbing Example:

Suppose that $y_t, t = 1, \ldots, n$, are drawings from a distribution $F(y)$. We want confidence intervals for some of the quantiles of $F(y)$.

- If q_α is the true α quantile and $\hat{q}_\alpha$ is the estimate, asymptotic theory tells us that

$$\mathrm{se}(\hat{q}_\alpha) = \left(\frac{\alpha(1-\alpha)}{n f^2(q_\alpha)} \right)^{1/2}. \tag{31}$$

  In practice, we replace $f(q_\alpha)$ by a kernel density estimate $\hat{f}(\hat{q}_\alpha)$.

- The 0.95 **asymptotic confidence interval** is

$$\left[\hat{q}_\alpha - 1.96\, \widehat{\mathrm{se}}(\hat{q}_\alpha),\ \ \hat{q}_\alpha + 1.96\, \widehat{\mathrm{se}}(\hat{q}_\alpha) \right]. \tag{32}$$

- Simplest bootstrap procedure is just to resample the data, calculate the quantile(s) of each bootstrap sample, and then use (18) to estimate the standard error.
- This yields the 0.95 **simple bootstrap interval**

$$\big[\hat{q}_\alpha - 1.96\,\mathrm{se}^*(\hat{q}_\alpha),\ \ \hat{q}_\alpha + 1.96\,\mathrm{se}^*(\hat{q}_\alpha)\big]. \tag{33}$$

- We can also use the percentile t method. This requires kernel estimation for the actual sample and for every bootstrap sample. A modified version of this uses bootstrap instead of asymptotic standard errors in (24).

In the experiments, $F(y)$ was $\chi^2(3)$, $B = 999$, and $\alpha = 0.1, 0.2, \ldots, 0.9$. The sample size n varied from 50 to 1600 by factors of $\sqrt{2}$.

Figures 3 and 4 show **coverage frequency** for the four intervals for the 0.5 quantile (the median) and the 0.9 quantile.

Coverage frequency is the proportion of the time that the true value of the quantile falls within the interval. Ideally, it should be 0.95.

Results are not in accord with standard bootstrap theory:

- Simple bootstrap interval, which is conceptually the easiest to calculate, clearly performs best for the 0.9 quantile.
- Asymptotic interval performs best for the median.
- Percentile t interval, which (inappropriate) theory recommends, is not very reliable.
- Modified percentile t interval that uses bootstrap standard errors performs even worse for the median than percentile t, but somewhat better for the 0.9 quantile when n is small.

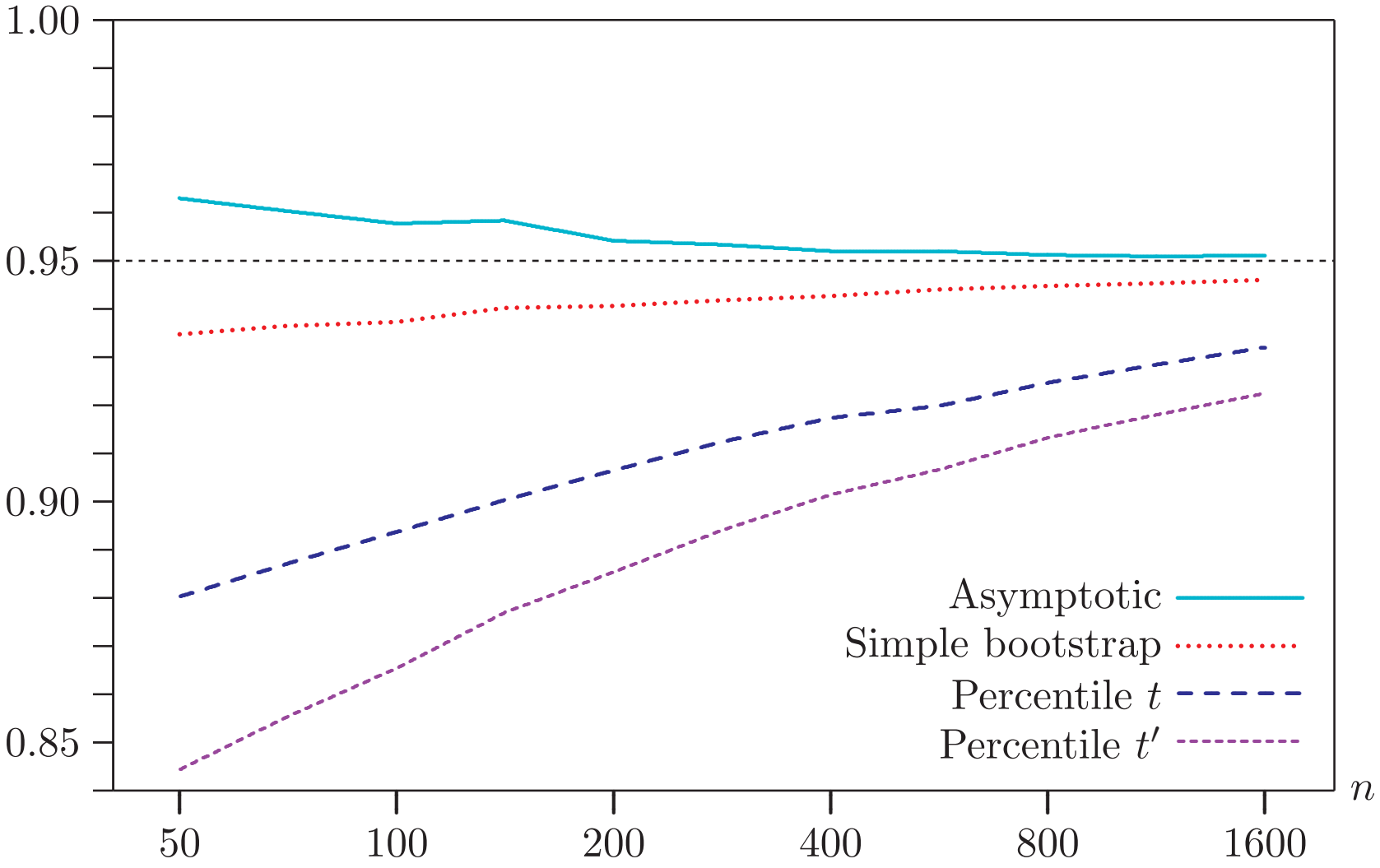


Figure 3. Coverage of four confidence intervals for the median

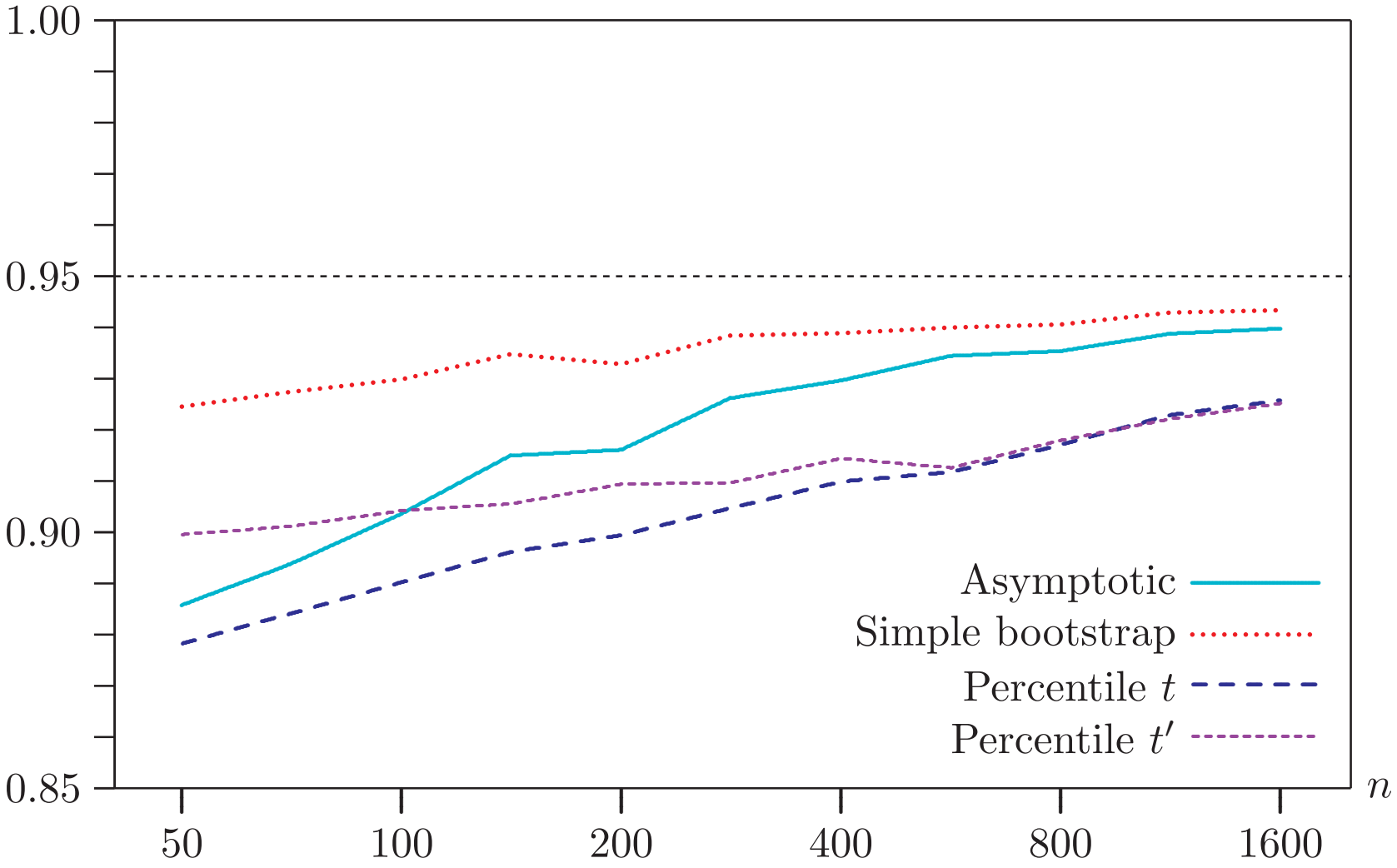


Figure 4. Coverage of four confidence intervals for the 0.9 quantile

Bootstrap DGPs for Dependent Data:

All bootstrap DGPs discussed so far assume that the errors are independent.

Resampling breaks up any dependence and is therefore inappropriate for dependent data.

For dependent data, there are two main approaches: **sieve bootstrap** and **block bootstrap**. Each has many variants.

1. The sieve bootstrap

Suppose that the error terms u_t in a regression model follow an unknown, stationary process with homoskedastic innovations.

The sieve bootstrap approximates this process using an AR(p) process, with p chosen by some sort of model selection criterion (like AIC or BIC), or by sequential testing.

1. Estimate the model (imposing the null hypothesis if one is to be tested) to obtain residuals $\hat{u}_t$.

2. Estimate AR(p) model

$$\hat{u}_t = \sum_{i=1}^{p} \rho_i \hat{u}_{t-i} + \varepsilon_t \tag{34}$$

for several values of p and choose best one. Can also use Yule-Walker equations so as to ensure stationarity.

3. Generate bootstrap error terms

$$u_t^* = \sum_{i=1}^{p} \hat{\rho}_i u_{t-i}^* + \varepsilon_t^*, \quad t = -99, \ldots, n, \tag{35}$$

where the ε_t^* are resampled from the (rescaled) residuals from (34). Set initial values of u_t^* to 0, and discard u_t^* for $t < 1$.

4. Generate the bootstrap data according to

$$y_t^* = \boldsymbol{X}_t \hat{\boldsymbol{\beta}} + u_t^*, \tag{36}$$

where $\hat{\boldsymbol{\beta}}$ may be restricted OLS, unrestricted OLS, restricted GLS, or unrestricted GLS estimates. GLS estimates would use covariance matrix implied by (35).

Sieve bootstrap assumes IID innovations, thus ruling out GARCH and other forms of heteroskedasticity.

The sieve bootstrap has recently been applied to Dickey-Fuller unit root testing by Park (2003) and Chang and Park (2003) .

2. Block bootstrap methods

Block bootstrap methods divide the quantities that are being resampled, which might be either rescaled residuals or $[\boldsymbol{y}, \boldsymbol{X}]$ pairs, into blocks of b consecutive observations. We then resample the blocks.

- Blocks may be either overlapping or nonoverlapping; overlapping seems to be better.

- Block lengths may be fixed or variable (as in **stationary bootstrap**); fixed seems to be better.
- For the **moving-block bootstrap**, there are $n - b + 1$ blocks. The first contains obs. 1 through b, the second contains obs. 2 through $b + 1$, and the last contains obs. $n - b + 1$ through n.
- Choice of b is critical. In theory, it must increase as n increases. Often proportional to $n^{1/3}$.
- If blocks are too short, bootstrap samples cannot mimic original sample. Dependence is broken whenever we start a new block.
- If blocks are too long, bootstrap samples are not random enough.
- The **block-of-blocks** bootstrap is the analog of the pairs bootstrap for dynamic models. Consider the dynamic regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \gamma y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \tag{37}$$

If we define

$$\boldsymbol{Z}_t \equiv [y_t, y_{t-1}, \boldsymbol{X}_t], \tag{38}$$

we can construct $n - b + 1$ overlapping blocks as

$$\boldsymbol{Z}_1 \dots \boldsymbol{Z}_b, \;\; \boldsymbol{Z}_2 \dots \boldsymbol{Z}_{b+1}, \;\; \dots\dots \;\; , \boldsymbol{Z}_{n-b+1} \dots \boldsymbol{Z}_n. \tag{39}$$

- The block-of-blocks bootstrap works with heteroskedasticity as well as serial correlation.
- Although block bootstrap methods frequently offer higher-order accuracy than asymptotic methods, they generally do so to only a modest extent. See Hall, Horowitz, and Jing (1995) and Andrews (2002, 2004).
- Block bootstrap can yield more reliable standard errors than using HAC covariance matrices; see Gonçalves and White (2005).

Recent surveys of bootstrap methods for time-series data include Bühlmann (2002), Politis (2003), and Härdle, Horowitz, and Kreiss (2003). See also Horowitz (2003). Gonçalves and White (2004) provide weak conditions for block bootstrap to be valid.

Example: Unit Root Tests

One version of the augmented Dickey-Fuller τ test is the t statistic for $\beta_1 = 0$ in the regression

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \sum_{j=1}^{p} \delta_j \Delta y_{t-j} + e_t, \tag{40}$$

where the p lags of Δy_{t-j} are added to account for serial correlation in the error terms. Can choose p in several ways, which substantially affect finite-sample properties (e.g. AIC, BIC, sequential testing).

To bootstrap this test, we need to run the regression under the null that $\beta_1 = 0$ and then generate bootstrap samples that satisfy the null.

1. For **moving-block bootstrap**, just calculate $\hat{u}_t = \Delta y_t - \Delta \bar{y}$. Then resample from blocks of the $\hat{u}_t$. Generate bootstrap data from

$$y_t^* = y_{t-1}^* + u_t^*. \tag{41}$$

   Note no constant term, since there is none under the null. Start with 0 and discard initial observations.

2. For **sieve bootstrap**, regress Δy_t on a constant and a number of lags of Δy_t, obtaining coefficients $\hat{\rho}_i$ and residuals $\hat{\varepsilon}_t$. Generate data from

$$y_t^* = y_{t-1}^* + \sum_{i=1}^{p} \hat{\rho}_i \Delta y_{t-i}^* + \varepsilon_t^*, \quad t = -l, \ldots, n, \tag{42}$$

using 0s for the initial values of y_{t-i}^*. The ε_t^* are resampled from rescaled $\hat{\varepsilon}_t$. There are several ways to choose p.

3. For **semiparametric bootstrap** assuming MA(1) errors, estimate the model

$$\Delta y_t = \beta_0 + \varepsilon_t + \alpha \varepsilon_{t-1}, \quad \varepsilon_t \sim \text{IID}(0, \sigma^2), \tag{43}$$

and generate bootstrap data from

$$y_t^* = y_{t-1}^* + \varepsilon_t^* + \hat{\alpha} \varepsilon_{t-1}^*, \tag{44}$$

where the ε_t^* are resampled from rescaled $\hat{\varepsilon}_t$.

Of course, this semiparametric bootstrap requires us to know that the errors are MA(1), which is highly implausible.

Experiments

- Errors were MA(1) with various values of α.
- $n = 50$
- $B = 399$
- 100,000 replications for 39 values of α.
- Between 4 and 12 lags in test regression, chosen by AIC.
- Block length 12 for moving-block bootstrap.
- Between 4 and 12 lags for sieve bootstrap, chosen by AIC.
- See Figure 5. All bootstrap methods work very well for $\alpha > 0$, but all work badly for $\alpha < -0.8$.

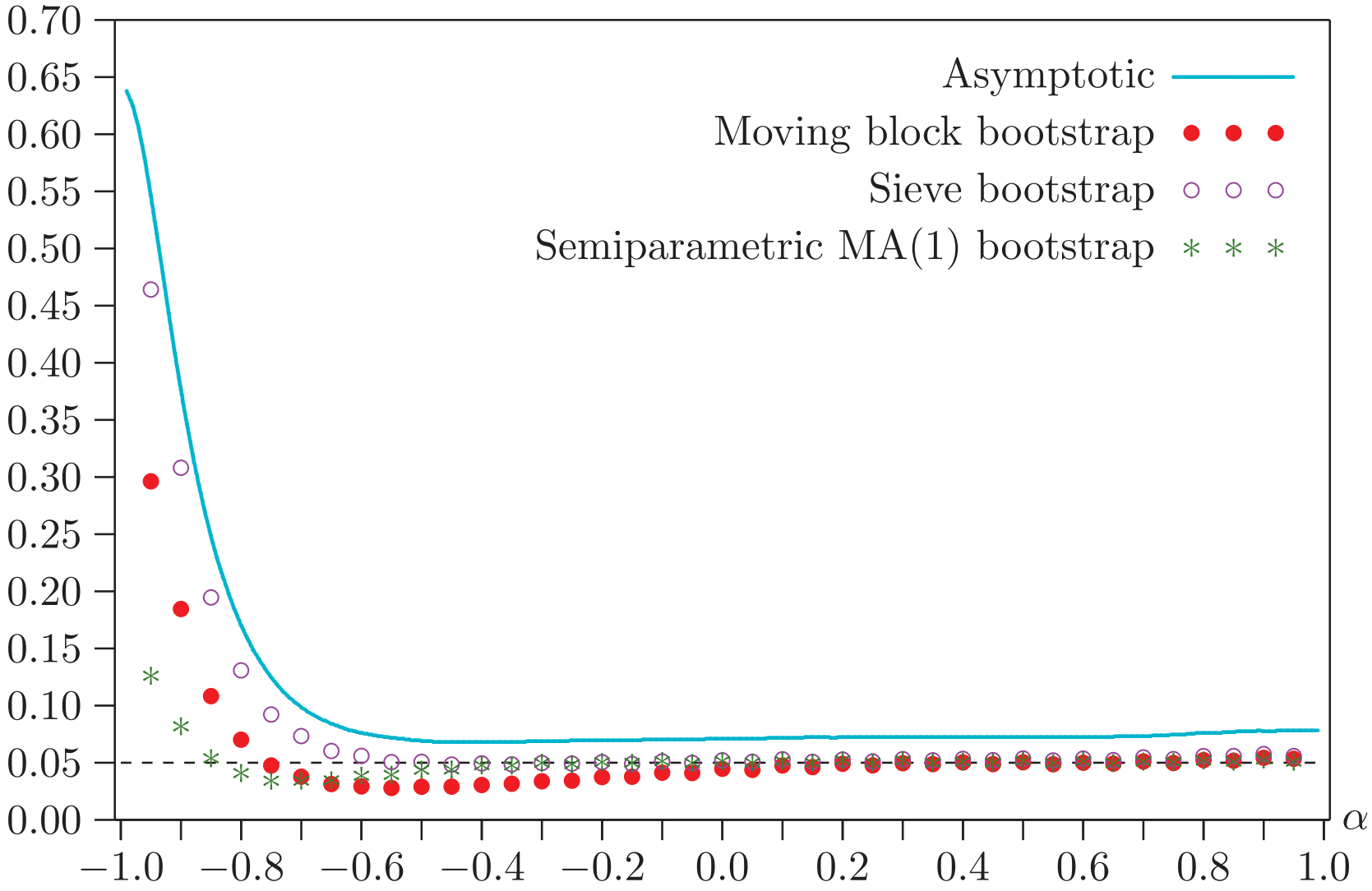


Figure 5. Rejection frequencies for Dickey-Fuller tests, $n = 50$

Conclusions

- Talking about "the bootstrap" is misleading. In fact, there are many bootstrap methods.
- The first, and often the hardest, thing to do is to choose the appropriate bootstrap DGP: restricted or unrestricted, parametric, residual, pairs, wild, moving-block, block of blocks, etc.
- The second thing to do is to decide how to use the bootstrap quantities to make inferences.
- For hypothesis testing, we must decide:
  - What statistic(s) to bootstrap and whether to impose the null hypothesis on the bootstrap DGP.
  - Whether to calculate bootstrap P value or bootstrap critical value; this should not affect outcome of the test.
  - For tests based on signed statistics, whether to assume symmetry or not.
  - Whether to use an asymptotic test based on a statistic that requires the bootstrap to compute.

- For confidence intervals, options include:
  - Asymptotic interval using bootstrap standard errors.
  - Bias-corrected interval using bootstrap standard errors.
  - Bootstrap t interval using asymptotic standard errors.
  - Bootstrap t interval using bootstrap standard errors.
  - Primitive methods like "naive" percentile interval.
  - More exotic methods like ABC and BC_a.
  - Impose symmetry or not.
  - Use nonlinear transformations or not.
  - Use unrestricted bootstrap DGP (usual approach) or better, but more complicated, methods that impose restrictions.

References

Andrews, D. W. K. (2002). "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators," *Econometrica*, 70, 119–162.

Andrews, D. W. K. (2004). "The block-block bootstrap: Improved asymptotic refinements," *Econometrica*, 72, 673–700.

Bühlmann, P. (2002). "Bootstraps for time series," *Statistical Science*, 17, 52–72.

Chang, Y., and J. Y. Park (2003). "A sieve bootstrap for the test of a unit root," *Journal of Time Series Analysis*, 24, 379–400.

Davidson, R., and E. Flachaire (2001). "The wild bootstrap, tamed at last," GREQAM Document de Travail 99A32, revised.

Davidson, R., and J. G. MacKinnon (1999). "The size distortion of bootstrap tests," *Econometric Theory*, 15, 361–376.

Davidson, R., and J. G. MacKinnon (2004). *Econometric Theory and Methods*, New York, Oxford University Press.

Davidson, R., and J. G. MacKinnon (2005a). "The power of bootstrap and asymptotic tests," *Journal of Econometrics*, forthcoming.

Davidson, R., and J. G. MacKinnon (2005b). "Bootstrap methods in econometrics," Chapter 23 in *Palgrave Handbook of Econometrics: Volume 1 Theoretical Econometrics*, ed. K. Patterson and T. C. Mills, Basingstoke, Palgrave Macmillan.

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

Dufour, J.-M., and L. Khalaf (2001). "Monte Carlo test methods in econometrics," Chapter 23 in *A Companion to Econometric Theory*, ed. B. Baltagi, Oxford, Blackwell Publishers, 494–519.

Dufour, J.-M., L. Khalaf, J.-T. Bernard, and I. Genest (2004). "Simulation-based finite-sample tests for heteroskedasticity and ARCH effects," *Journal of Econometrics*, 122, 317–347.

Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses," *Annals of Mathematical Statistics*, 28, 181–187.

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia, Society for Industrial and Applied Mathematics.

Freedman, D. A. (1981). "Bootstrapping regression models," *Annals of Statistics*, 9, 1218–1228.

Freedman, D. A. (1984). "On bootstrapping stationary two-stage least-squares estimates in stationary linear models," *Annals of Statistics*, 12, 827–842.

Freedman, D. A., and S. C. Peters (1984). "Bootstrapping an econometric model: Some empirical results," *Journal of Business and Economic Statistics*, 2, 150–158.

Gonçalves, S., and L. Kilian (2004). "Bootstrapping autoregressions with heteroskedasticity of unknown form," *Journal of Econometrics*, 123, 89–120.

Gonçalves, S., and H. White (2004). "Maximum likelihood and the bootstrap for dynamic nonlinear models," *Journal of Econometrics*, 119, 199–219.

Gonçalves, S., and H. White (2005). "Bootstrap standard errors for linear regression" *Journal of the American Statistical Association*, forthcoming.

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

Hall, P., J. L. Horowitz, and B. Y. Jing (1995). "On blocking rules for the bootstrap with dependent data," *Biometrika*, 82, 561–574.

Hansen, B. E. (2000). "Testing for structual change in conditional models," *Journal of Econometrics*, 97, 93–115.

Härdle, W., J. L. Horowitz, and J.-P. Kreiss (2003). "Bootstrap methods for time series," *International Statistical Review*, 71, 435–459.

Horowitz, J. L. (2003). "The bootstrap in econometrics," *Statistical Science*, 18, 211–218.

Horowitz, J. L., and N. E. Savin (2000). "Empirically relevant critical values for hypothesis tests," *Journal of Econometrics*, 95, 375–389.

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

MacKinnon, J. G. (2002). "Bootstrap inference in econometrics," *Canadian Journal of Economics*, 35 615–645.

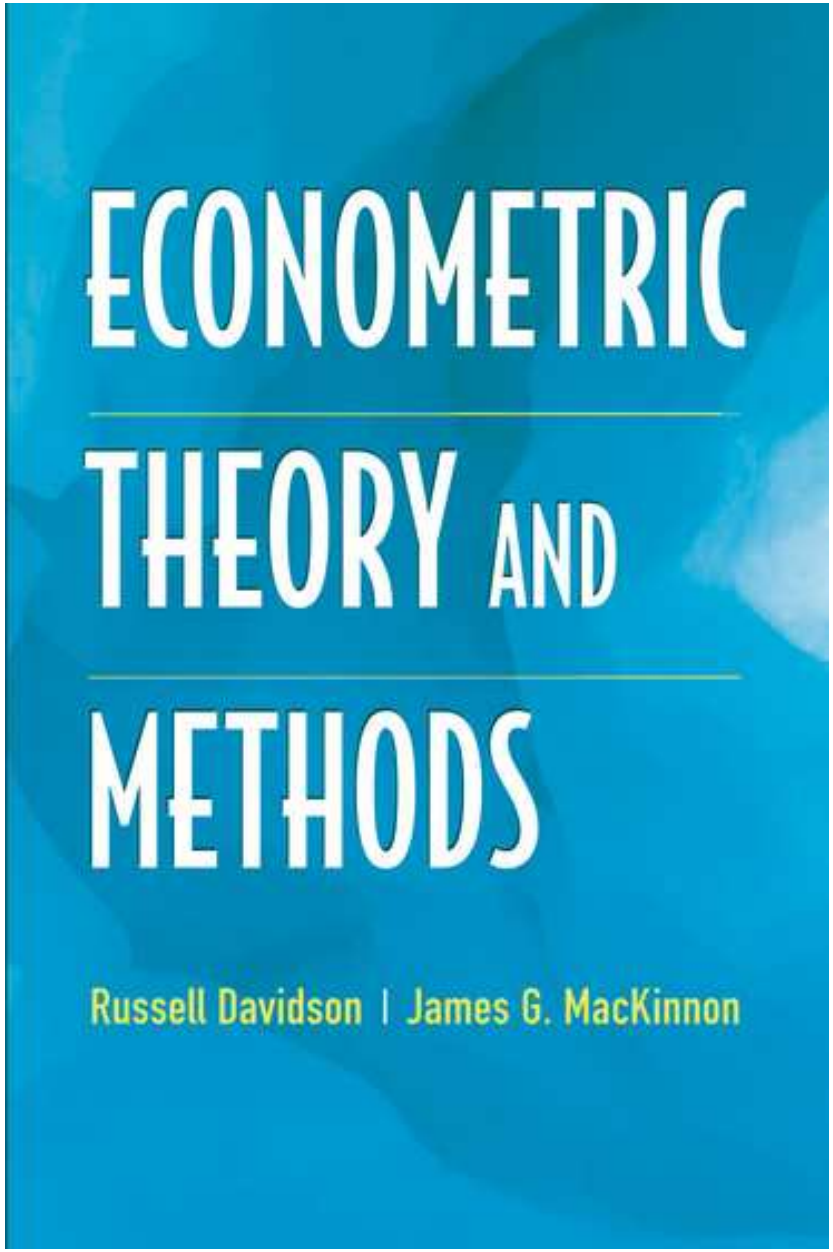
MacKinnon, J. G., and Anthony A. Smith, Jr. (1998). "Approximate bias correction in econometrics," *Journal of Econometrics*, 85 205–230.

Park, J. Y. (2003). "Bootstrap unit root tests," *Econometrica*, 71, 1845–1895.

Politis, D. N. (2003). "The impact of bootstrap methods on time series analysis," *Statistical Science*, 18, 219–230.

Racine, J. and J. G. MacKinnon (2004). "Simulation-based tests that can use any number of simulations," manuscript.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis," *Annals of Statistics*, 14, 1261–1295.


ECONOMETRIC
THEORY AND
METHODS
Russell Davidson | James G. MacKinnon