CHAPTER 11

Bootstrap Methods

11.1. Introduction

Exact finite-sample results are unavailable for most microeconometrics estimators and related test statistics. The statistical inference methods presented in preceding chapters rely on asymptotic theory that usually leads to limit normal and chi-square distributions.

An alternative approximation is provided by the bootstrap, due to Efron (1979, 1982). This approximates the distribution of a statistic by a Monte Carlo simulation, with sampling done from the empirical distribution or the fitted distribution of the observed data. The additional computation required is usually feasible given advances in computing power. Like conventional methods, however, bootstrap methods rely on asymptotic theory and are only exact in infinitely large samples.

The wide range of bootstrap methods can be classified into two broad approaches. First, the simplest bootstrap methods can permit statistical inference when conventional methods such as standard error computation are difficult to implement. Second, more complicated bootstraps can have the additional advantage of providing asymptotic refinements that can lead to a better approximation in-finite samples. Applied researchers are most often interested in the first aspect of the bootstrap. Theoreticians emphasize the second, especially in settings where the usual asymptotic methods work poorly in finite samples.

The econometrics literature focuses on use of the bootstrap in hypothesis testing, which relies on approximation of probabilities in the tails of the distributions of statistics. Other applications are to confidence intervals, estimation of standard errors, and bias reduction. The bootstrap is straightforward to implement for smooth $\sqrt{N}$-consistent estimators based on iid samples, though bootstraps with asymptotic refinements are underutilized. Caution is needed in other settings, including nonsmooth estimators such as the median, nonparametric estimators, and inference for data that are not iid.

A reasonably self-contained summary of the bootstrap is provided in Section 11.2, an example is given in Section 11.3, and some theory is provided in Section 11.4.

Further variations of the bootstrap are presented in Section 11.5. Section 11.6 presents use of the bootstrap for specific types of data and specific methods used often in microeconometrics.

11.2. Bootstrap Summary

We summarize key bootstrap methods for estimator $\widehat{\boldsymbol{\theta}}$ and associated statistics based on an iid sample $\{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$, where usually $\mathbf{w}_i = (y_i, \mathbf{x}_i)$ and $\widehat{\boldsymbol{\theta}}$ is a smooth estimator that is $\sqrt{N}$ consistent and asymptotically normally distributed. For notational simplicity we generally present results for scalar θ. For vector $\boldsymbol{\theta}$ in most instances replace θ by θ_j, the jth component of $\boldsymbol{\theta}$.

Statistics of interest include the usual regression output: the estimate $\widehat{\theta}$; standard errors $s_{\widehat{\theta}}$; t-statistic $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$, where θ_0 is the null hypothesis value; the associated critical value or p-value for this statistic; and a confidence interval.

This section presents bootstraps for each of these statistics. Some motivation is also provided, with the underlying theory sketched in Section 11.4.

11.2.1. Bootstrap without Refinement

Consider estimation of the variance of the sample mean $\widehat{\mu} = \bar{y} = N^{-1}\sum_{i=1}^{N} y_i$, where the scalar random variable y_i is iid $[\mu, \sigma^2]$, when it is not known that $\mathrm{V}[\widehat{\mu}] = \sigma^2/N$.

The variance of $\widehat{\mu}$ could be obtained by obtaining S such samples of size N from the population, leading to S sample means and hence S estimates $\widehat{\mu}_s = \bar{y}_s$, $s = 1, \ldots, S$. Then we could estimate $\mathrm{V}[\widehat{\mu}]$ by $(S-1)^{-1}\sum_{s=1}^{S}(\widehat{\mu}_s - \overline{\widehat{\mu}})^2$, where $\overline{\widehat{\mu}} = S^{-1}\sum_{s=1}^{S}\widehat{\mu}_s$.

Of course this approach is not possible, as we only have one sample. A bootstrap can implement this approach by viewing the sample as the population. Then the finite population is now the actual data $y_1, \ldots, y_N$. The distribution of $\widehat{\mu}$ can be obtained by drawing B bootstrap samples from this population of size N, where each bootstrap sample of size N is obtained by sampling from $y_1, \ldots, y_N$ with replacement. This leads to B sample means and hence B estimates $\widehat{\mu}_b = \bar{y}_b$, $b = 1, \ldots, B$. Then estimate $\mathrm{V}[\widehat{\mu}]$ by $(B-1)^{-1}\sum_{b=1}^{B}(\widehat{\mu}_b - \overline{\widehat{\mu}})^2$, where $\overline{\widehat{\mu}} = B^{-1}\sum_{b=1}^{B}\widehat{\mu}_b$. Sampling with replacement may seem to be a departure from usual sampling methods, but in fact standard sampling theory assumes sampling with replacement rather than without replacement (see Section 24.2.2).

With additional information other ways to obtain bootstrap samples may be possible. For example, if it is known that $y_i \sim \mathcal{N}[\mu, \sigma^2]$ then we could obtain B bootstrap samples of size N by drawing from the $\mathcal{N}[\widehat{\mu}, s^2]$ distribution. This bootstrap is an example of a parametric bootstrap, whereas the preceding bootstrap was from the empirical distribution.

More generally, for estimator $\widehat{\boldsymbol{\theta}}$ similar bootstraps can be used to, for example, estimate $\mathrm{V}[\widehat{\boldsymbol{\theta}}]$ and hence standard errors when analytical formulas for $\mathrm{V}[\widehat{\boldsymbol{\theta}}]$ are complex. Such bootstraps are usually valid for observations $\mathbf{w}_i$ that are iid over i, and they have similar properties to estimates obtained using the usual asymptotic theory.

11.2.2. Asymptotic Refinements

In some settings it is possible to improve on the preceding bootstrap and obtain estimates that are equivalent to those obtained using a more refined asymptotic theory that may better approximate the finite-sample distribution of $\widehat{\boldsymbol{\theta}}$. Much of this chapter is directed to such **asymptotic refinements**.

Usual asymptotic theory uses the result that $\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[0, \sigma^2]$. Thus

$$\Pr[\sqrt{N}(\widehat{\theta} - \theta_0)/\sigma \leq z] = \Phi(z) + R_1, \tag{11.1}$$

where $\Phi(\cdot)$ is the standard normal cdf and R_1 is a remainder term that disappears as $N \to \infty$.

This result is based on asymptotic theory detailed in Section 5.3 that includes application of a central limit theorem. The CLT is based on a truncated power-series expansion. The Edgeworth expansion, detailed in Section 11.4.3, includes additional terms in the expansion. With one extra term this yields

$$\Pr[\sqrt{N}(\widehat{\theta} - \theta_0)/\sigma \leq z] = \Phi(z) + \frac{g_1(z)\phi(z)}{\sqrt{N}} + R_2, \tag{11.2}$$

where $\phi(\cdot)$ is the standard normal density, $g_1(\cdot)$ is a bounded function given after (11.13) in Section 11.4.3 and R_2 is a remainder term that disappears as $N \to \infty$.

The Edgeworth expansion is difficult to implement theoretically as the function $g_1(\cdot)$ is data dependent in a complicated way. A bootstrap *with asymptotic refinement* provides a simple computational method to implement the Edgeworth expansion. The theory is given in Section 11.4.4.

Since $R_1 = O(N^{-1/2})$ and $R_2 = O(N^{-1})$, asymptotically $R_2 < R_1$, leading to a better approximation as $N \to \infty$. However, in finite samples it is possible that $R_2 > R_1$. A bootstrap with asymptotic refinement provides a better approximation asymptotically that hopefully leads to a better approximation in samples of the finite sizes typically used. Nevertheless, there is no guarantee and simulation studies are frequently used to verify that finite-sample gains do indeed occur.

11.2.3. Asymptotically Pivotal Statistic

For asymptotic refinement to occur, the statistic being bootstrapped must be an **asymptotically pivotal statistic**, meaning a statistic whose limit distribution does not depend on unknown parameters. This result is explained in Section 11.4.4.

As an example, consider sampling from $y_i \sim [\mu, \sigma^2]$. Then the estimate $\widehat{\mu} = \bar{y} \overset{a}{\sim} \mathcal{N}[\mu, \sigma^2/N]$ is not asymptotically pivotal even given a null hypothesis value $\mu = \mu_0$ since its distribution depends on the unknown parameter σ^2. However, the **studentized statistic** $t = (\widehat{\mu} - \mu_0)/s_{\widehat{\mu}} \overset{a}{\sim} \mathcal{N}[0, 1]$ is asymptotically pivotal.

Estimators are usually not asymptotically pivotal. However, conventional asymptotically standard normal or chi-squared distributed test statistics, including Wald, Lagrange multiplier, and likelihood ratio tests, and related confidence intervals, are asymptotically pivotal.

11.2.4. The Bootstrap

In this section we provide a broad description of the bootstrap, with further details given in subsequent sections.

Bootstrap Algorithm

A general **bootstrap algorithm** is as follows:

1. Given data $\mathbf{w}_1, \ldots, \mathbf{w}_N$, draw a bootstrap sample of size N using a method given in the following and denote this new sample $\mathbf{w}_1^*, \ldots, \mathbf{w}_N^*$.
2. Calculate an appropriate statistic using the bootstrap sample. Examples include (a) the estimate $\widehat{\theta}^*$ of θ, (b) the standard error $s_{\widehat{\theta}^*}$ of the estimate $\widehat{\theta}^*$, and (c) a t-statistic $t^* = (\widehat{\theta}^* - \widehat{\theta})/s_{\widehat{\theta}^*}$ centered at the original estimate $\widehat{\theta}$. Here $\widehat{\theta}^*$ and $s_{\widehat{\theta}^*}$ are calculated in the usual way but using the new bootstrap sample rather than the original sample.
3. Repeat steps 1 and 2 B independent times, where B is a large number, obtaining B bootstrap replications of the statistic of interest, such as $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ or $t_1^*, \ldots, t_B^*$.
4. Use these B bootstrap replications to obtain a bootstrapped version of the statistic, as detailed in the following subsections.

Implementation can vary according to how bootstrap samples are obtained, how many bootstraps are performed, what statistic is being bootstrapped, and whether or not that statistic is asymptotically pivotal.

Bootstrap Sampling Methods

The bootstrap dgp in step 1 is used to approximate the true unknown dgp.

The simplest bootstrapping method is to use the empirical distribution of the data, which treats the sample as being the population. Then $\mathbf{w}_1^*, \ldots, \mathbf{w}_N^*$ are obtained by sampling with replacement from $\mathbf{w}_1, \ldots, \mathbf{w}_N$. In each bootstrap sample so obtained, some of the original data points will appear multiple times whereas others will not appear at all. This method is an **empirical distribution function (EDF) bootstrap** or **nonparametric bootstrap**. It is also called a **paired bootstrap** since in single-equation regression models $\mathbf{w}_i = (y_i, \mathbf{x}_i)$, so here both y_i and $\mathbf{x}_i$ are resampled.

Suppose the conditional distribution of the data is specified, say $y|\mathbf{x} \sim F(\mathbf{x}, \boldsymbol{\theta}_0)$, and an estimate $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ is available. Then in step 1 we can instead form a bootstrap sample by using the original $\mathbf{x}_i$ while generating y_i by random draws from $F(\mathbf{x}_i, \widehat{\boldsymbol{\theta}})$. This corresponds to regressors fixed in repeated samples (see Section 4.4.5). Alternatively, we may first resample $\mathbf{x}_i^*$ from $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and then generate y_i from $F(\mathbf{x}_i^*, \widehat{\boldsymbol{\theta}})$, $i = 1, \ldots, N$. Both are examples of a **parametric bootstrap** that can be applied in fully parametric models.

For regression model with additive iid error, say $y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + u_i$, we can form fitted residuals $\widehat{u}_1, \ldots, \widehat{u}_N$, where $\widehat{u}_i = y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})$. Then in step 1 bootstrap from these residuals to get a new draw of residuals, say $(\widehat{u}_1^*, \ldots, \widehat{u}_N^*)$, leading to a bootstrap sample $(y_1^*, \mathbf{x}_1), \ldots, (y_N^*, \mathbf{x}_N)$, where $y_i^* = g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}) + u_i^*$. This bootstrap is called a

residual bootstrap. It uses information intermediate between the nonparametric and parametric bootstrap. It can be applied if the error term has distribution that does not depend on unknown parameters.

We emphasize the paired bootstrap on grounds of its simplicity, applicability to a wide range of nonlinear models, and reliance on weak distributional assumptions. However, the other bootstraps generally provide a better approximation (see Horowitz, 2001, p. 3185) and should be used if the stronger model assumptions they entail are warranted.

The Number of Bootstraps

The bootstrap asymptotics rely on $N \rightarrow \infty$ and so the bootstrap can be asymptotically valid even for low B. However, clearly the bootstrap is more accurate as $B \rightarrow \infty$. A sufficiently large value of B varies with one's tolerance for bootstrap-induced simulation error and with the purpose of the bootstrap.

Andrews and Buchinsky (2000) present an application-specific numerical method to determine the number of replications B needed to ensure a given level of accuracy or, equivalently, the level of accuracy obtained for a given value of B. Let λ denote the quantity of interest, such as a standard error or a critical value, $\widehat{\lambda}_\infty$ denote the ideal bootstrap estimate with $B = \infty$, and $\widehat{\lambda}_B$ denote the estimate with B bootstraps. Then Andrews and Buchinsky (2000) show that

$$\sqrt{B}(\widehat{\lambda}_B - \widehat{\lambda}_\infty)/\widehat{\lambda}_\infty \xrightarrow{d} \mathcal{N}[0, \omega],$$

where ω varies with the application and is defined in Table III of Andrews and Buchinsky (2000). It follows that $\Pr[\delta \leq z_{\tau/2}\sqrt{\omega/B}] = 1 - \tau$, where $\delta = |\widehat{\lambda}_B - \widehat{\lambda}_\infty|/\widehat{\lambda}_\infty$ denotes the relative discrepancy caused by only B replications. Thus $B \geq \omega z^2_{\tau/2}/\delta^2$ ensures the relative discrepancy is less than δ with probability at least $1 - \tau$. Alternatively, given B replications the relative discrepancy is less than $\delta = z_{\tau/2}\sqrt{\omega/B}$.

To provide concrete guidelines we propose the rule of thumb that

$$B = 384\omega.$$

This ensures that the relative discrepancy is less than 10% with probability at least 0.95, since $z^2_{.025}/0.1^2 = 384$. The only difficult part in implementation is estimation of ω, which varies with the application.

For standard error estimation $\omega = (2 + \gamma_4)/4$, where γ_4 is the coefficient of excess kurtosis for the bootstrap estimator $\widehat{\theta}^*$. Intuitively, fatter tails in the distribution of the estimator mean outliers are more likely, contaminating standard error estimation. It follows that $B = 384 \times (1/2) = 192$ is enough if $\gamma_4 = 0$ whereas $B = 960$ is needed if $\gamma_4 = 8$. These values are higher than those proposed by Efron and Tibsharani (1993, p. 52), who state that $B = 200$ is almost always enough.

For a symmetric two-sided test or confidence interval at level α, $\omega = \alpha(1 - \alpha)/[2z_{\alpha/2}\phi(z_{\alpha/2})]^2$. This leads to $B = 348$ for $\alpha = 0.05$ and $B = 685$ for $\alpha = 0.01$. As expected more bootstraps are needed the further one goes into the tails of the distribution.

For a one-sided test or nonsymmetric two-sided test or confidence interval at level α, $\omega = \alpha(1-\alpha)/[z_\alpha \phi(z_\alpha)]^2$. This leads to $B = 634$ for $\alpha = 0.05$ and $B = 989$ for $\alpha = 0.01$. More bootstraps are needed when testing in one tail. For chi-squared tests with h degrees of freedom $\omega = \alpha(1-\alpha)/[\chi^2_\alpha(h) f(\chi^2_\alpha(h))]^2$, where $f(\cdot)$ is the $\chi^2(h)$ density.

For test p-values $\omega = (1-p)/p$. For example, if $p = 0.05$ then $\omega = 19$ and $B =$ 7,296. Many more bootstraps are needed for precise calculation of the test p-value compared to hypothesis rejection if a critical value is exceeded.

For bias-corrected estimation of θ a simple rule uses $\widehat{\omega} = \widehat{\sigma}^2/\widehat{\theta}^2$, where the estimator $\widehat{\theta}$ has standard error $\widehat{\sigma}$. For example, if the usual t-statistic $t = \widehat{\theta}/\widehat{\sigma} = 2$ then $\widehat{\omega} = 1/4$ and $B = 96$. Andrews and Buchinsky (2000) provide many more details and refinements of these results.

For hypothesis testing, Davidson and MacKinnon (2000) provide an alternative approach. They focus on the loss of power caused by bootstrapping with finite B. (Note that there is no power loss if $B = \infty$.) On the basis of simulations they recommend at least $B = 399$ for tests at level 0.05, and at least $B = 1{,}499$ for tests at level 0.01. They argue that for testing their approach is superior to that of Andrews and Buchinsky.

Several other papers by Davidson and MacKinnon, summarized in MacKinnon (2002), emphasize practical considerations in bootstrap inference. For hypothesis testing at level α choose B so that $\alpha(B+1)$ is an integer. For example, at $\alpha = 0.05$ let $B = 399$ rather than 400. If instead $B = 400$ it is unclear on an upper one-sided alternative test whether the 20th or 21st largest bootstrap t-statistic is the critical value. For nonlinear models computation can be reduced by performing only a few Newton–Raphson iterations in each bootstrap sample from starting values equal to the initial parameter estimates.

11.2.5. Standard Error Estimation

The **bootstrap estimate of variance** of an estimator is the usual formula for estimating a variance, applied to the B bootstrap replications $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$:

$$s^2_{\widehat{\theta},\text{Boot}} = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}^*_b - \overline{\widehat{\theta}}^*)^2, \tag{11.3}$$

where

$$\overline{\widehat{\theta}}^* = B^{-1}\sum_{b=1}^{B}\widehat{\theta}^*_b. \tag{11.4}$$

Taking the square root yields $s_{\widehat{\theta},\text{Boot}}$, the **bootstrap estimate of the standard error**.

This bootstrap provides no asymptotic refinement. Nonetheless, it can be extraordinarily useful when it is difficult to obtain standard errors using conventional methods. There are many examples. The estimate $\widehat{\theta}$ may be a **sequential two-step m-estimator** whose standard error is difficult to compute using the results given in Secttion 6.8. The estimate $\widehat{\theta}$ may be a 2SLS estimator estimated using a package that

only reports standard errors assuming homoskedastic errors but the errors are actually **heteroskedastic**. The estimate $\widehat{\theta}$ may be a **function of other parameters** that are actually estimated, for example, $\widehat{\theta} = \widehat{\alpha}/\widehat{\beta}$, and the bootstrap can be used instead of the delta method. For **clustered data** with many small clusters, such as short panels, cluster-robust standard errors can be obtained by resampling the clusters.

Since the bootstrap estimate $s_{\widehat{\theta},\text{Boot}}$ is consistent, it can be used in place of $s_{\widehat{\theta}}$ in the usual asymptotic formula to form confidence intervals and hypothesis tests that are asymptotically valid. Thus asymptotic statistical inference is possible in settings where it is difficult to obtain standard errors by other methods. However, there will be **no improvement** in finite-sample performance. To obtain an asymptotic refinement the methods of the next section are needed.

11.2.6. Hypothesis Testing

Here we consider tests on an individual coefficient, denoted θ. The test may be either an upper one-tailed alternative of $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$ or a two-sided test of $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$. Other tests are deferred to Section 11.6.3.

Tests with Asymptotic Refinement

The usual test statistic $T_N = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$ provides the potential for asymptotic refinement, as it is asymptotically pivotal since its asymptotic standard normal distribution does not depend on unknown parameters. We perform B bootstrap replications producing B test statistics $t_1^*, \ldots, t_B^*$, where

$$t_b^* = (\widehat{\theta}_b^* - \widehat{\theta})/s_{\widehat{\theta}_b^*}. \tag{11.5}$$

The estimates t_b^* are centered around the original estimate $\widehat{\theta}$ since resampling is from a distribution centered around $\widehat{\theta}$. The empirical distribution of $t_1^*, \ldots, t_B^*$, ordered from smallest to largest, is then used to approximate the distribution of T_N as follows.

For an upper one-tailed alternative test the **bootstrap critical value** (at level α) is the upper α quantile of the B ordered test statistics. For example, if $B = 999$ and $\alpha = 0.05$ then the critical value is the 950th highest value of t^*, since then $(B + 1)(1 - \alpha) = 950$. For a similar lower tail one-sided test the critical value is the 50th smallest value of t^*.

One can also compute a **bootstrap** p**-value** in the obvious way. For example, if the original statstistic t lies between the 914th and 915th largest values of 999 bootstrap replicates then the p-value for a upper one-tailed alternative test is $1 - 914/(B + 1) = 0.086$.

For a two-sided test a distinction needs to be made between symmetrical and nonsymmetrical tests. For a **nonsymmetrical test** or **equal-tailed test** the bootstrap **critical values** (at level α) are the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the ordered test statistics t^*, and the null hypothesis is rejected at level α if the original t-statistic lies outside this range. For a **symmetrical test** we instead order $|t^*|$ and the bootstrap

critical value (at level α) is the upper α quantile of the ordered $|t^*|$. The null hypothesis is rejected at level α if $|t|$ exceeds this critical value.

These tests, using the **percentile-t method**, provide asymptotic refinements. For a one-sided t-test and for a nonsymmetrical two-sided t-test the true size of the test is $\alpha + O(N^{-1/2})$ with standard asymptotic critical values and $\alpha + O(N^{-1})$ with bootstrap critical values. For a two-sided symmetrical t-test or for an asymptotic chi-square test the asymptotic approximations work better, and the true size of the test is $\alpha + O(N^{-1})$ using standard asymptotic critical values and $\alpha + O(N^{-2})$ using bootstrap critical values.

Tests without Asymptotic Refinement

Alternative bootstrap methods can be used that although asymptotically valid do not provide an asymptotic refinement.

One approach already mentioned at the end of Section 11.2.5 is to compute $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta},\text{boot}}$, where the bootstrap estimate $s_{\widehat{\theta},\text{boot}}$ given in (11.3) replaces the usual estimate $s_{\widehat{\theta}}$, and compare this test statistic to critical values from the standard normal distribution.

A second approach, exposited here for a two-sided test of $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$, finds the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the bootstrap estimates $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$ and rejects H_0 if θ_0 falls outside this region. This is called the **percentile method**. Asymptotic refinement is obtained by using t^*_b in (11.5) that centers around $\widehat{\theta}$ rather than θ_0 and using a different standard error $s^*_{\widehat{\theta}}$ in each bootstrap.

These two bootstraps have the attraction of not requiring computation of $s_{\widehat{\theta}}$, the usual standard error estimate based on asymptotic theory.

11.2.7. Confidence Intervals

Much of the statistics literature considers confidence interval estimation rather than its flip side of hypothesis tests. Here instead we began with hypothesis tests, so only a brief presentation of confidence intervals is necessary.

An asymptotic refinement is based on the t-statistic, which is asymptotically pivotal. Thus from steps 1–3 in Section 11.2.4 we obtain bootstrap replication t-statistics $t^*_1, \ldots, t^*_B$. Then let $t^*_{[1-\alpha/2]}$ and $t^*_{[\alpha/2]}$ denote the lower and upper $\alpha/2$ quantiles of these t-statistics. The **percentile-t method** $100(1-\alpha)$ percent confidence interval is

$$\left(\widehat{\theta} - t^*_{[1-\alpha/2]} \times s_{\widehat{\theta}}, \widehat{\theta} + t^*_{[\alpha/2]} \times s_{\widehat{\theta}}\right), \tag{11.6}$$

where $\widehat{\theta}$ and $s_{\widehat{\theta}}$ are the estimate and standard error from the original sample.

An alternative is the **bias-corrected and accelerated (BC_a) method** detailed in Efron (1987). This offers an asymptotic refinement in a wider class of problems than the percentile-t method.

Other methods provide an asymptotically valid confidence interval, but without asymptotic refinement. First, one can use the bootstrap estimate of the standard

error in the usual confidence interval formula, leading to interval $(\widehat{\theta} - z_{[1-\alpha/2]} \times s_{\widehat{\theta},\text{boot}}, \widehat{\theta} + z_{[\alpha/2]} \times s_{\widehat{\theta},\text{boot}})$. Second, the **percentile method** confidence interval is the distance between the lower $\alpha/2$ and upper $\alpha/2$ quantiles of the B bootstrap estimates $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$ of θ.

11.2.8. Bias Reduction

Nonlinear estimators are usually biased in finite samples, though this bias goes to zero asymptotically if the estimator is consistent. For example, if μ^3 is estimated by $\widehat{\theta} = \bar{y}^3$, where y_i is iid $[\mu, \sigma^2]$, then $\text{E}[\widehat{\theta} - \mu^3] = 3\mu\sigma^2/N + \text{E}[(y-\mu)^3]/N^2$.

More generally, for a $\sqrt{N}$-consistent estimator

$$\text{E}[\widehat{\theta} - \theta_0] = \frac{a_N}{N} + \frac{b_N}{N^2} + \frac{c_N}{N^3} + \ldots, \tag{11.7}$$

where a_N, b_N, and c_N are bounded constants that vary with the data and estimator (see Hall, 1992, p. 53). An alternative estimator $\widetilde{\theta}$ provides an asymptotic refinement if

$$\text{E}[\widetilde{\theta} - \theta_0] = \frac{B_N}{N^2} + \frac{C_N}{N^3} + \ldots, \tag{11.8}$$

where B_N and C_N are bounded constants. For both estimators the bias disappears as $N \to \infty$. The latter estimator has the attraction that the bias goes to zero at a faster rate, and hence it is an asymptotic refinement, though in finite samples it is possible that $(B_N/N^2) > (a_N/N + b_N/N^2)$.

We wish to estimate the bias $\text{E}[\widehat{\theta}] - \theta$. This is the distance between the expected value or population average value of the parameter and the parameter value generating the data. The bootstrap replaces the population with the sample, so that the bootstrap samples are generated by parameter $\widehat{\theta}$, which has average value $\overline{\widehat{\theta}}^*$ over the bootstraps. The **bootstrap estimate of the bias** is then

$$\text{Bias}_{\widehat{\theta}} = (\overline{\widehat{\theta}}^* - \widehat{\theta}), \tag{11.9}$$

where $\overline{\widehat{\theta}}^*$ is defined in (11.4).

Suppose, for example, that $\widehat{\theta} = 4$ and $\overline{\widehat{\theta}}^* = 5$. Then the estimated bias is $(5-4) = 1$, an upward bias of 1. Since $\widehat{\theta}$ overestimates by 1, bias correction requires *subtracting* 1 from $\widehat{\theta}$, giving a bias-corrected estimate of 3. More generally, the **bootstrap bias-corrected estimator** of θ is

$$\begin{aligned} \widehat{\theta}_{\text{Boot}} &= \widehat{\theta} - (\overline{\widehat{\theta}}^* - \widehat{\theta}) \\ &= 2\widehat{\theta} - \overline{\widehat{\theta}}^*. \end{aligned} \tag{11.10}$$

Note that $\overline{\widehat{\theta}}^*$ itself is not the bias-corrected estimate. For more details on the direction of the correction, which may seem puzzling, see Efron and Tibsharani (1993, p. 138). For typical $\sqrt{N}$-consistent estimators the asymptotic bias of $\widehat{\theta}$ is $O(N^{-1})$ whereas the asymptotic bias of $\widehat{\theta}_{\text{Boot}}$ is instead $O(N^{-2})$.

In practice bias correction is seldom used for $\sqrt{N}$-consistent estimators, as the bootstrap estimate can be more variable than the original estimate $\widehat{\theta}$ and the bias is often

small relative to the standard error of the estimate. Bootstrap bias correction is used for estimators that converge at rate less than $\sqrt{N}$, notably nonparametric regression and density estimators.

11.3. Bootstrap Example

As a bootstrap example, consider the exponential regression model introduced in Section 5.9. Here the data are generated from an exponential distribution with an exponential mean with two regressors:

$$\begin{gathered} y_i|\mathbf{x}_i \sim \text{exponential}(\lambda_i),\ i = 1, \ldots, 50, \\ \lambda_i = \exp(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i}), \\ (x_{2i}, x_{3i}) \sim \mathcal{N}[0.1, 0.1; 0.1^2, 0.1^2, 0.005], \\ (\beta_1, \beta_2, \beta_3) = (-2, 2, 2). \end{gathered}$$

Maximum likelihood estimation on a sample of 50 observations yields $\widehat{\beta}_1 = -2.192$; $\widehat{\beta}_2 = 0.267$, $s_2 = 1.417$, and $t_2 = 0.188$; and $\widehat{\beta}_3 = 4.664$, $s_3 = 1.741$, and $t_3 = 2.679$. For this ML example the standard errors were based on $-\widehat{\mathbf{A}}^{-1}$, minus the inverse of the estimated Hessian matrix.

We concentrate on statistical inference for β_3 and demonstrate the bootstrap for standard error computation, test of statistical significance, confidence intervals, and bias correction. The differences between bootstrap and usual asymptotic estimates are relatively small in this example and can be much larger in other examples.

The results reported here are based on the paired bootstrap (see Section 11.2.4) with (y_i, x_{2i}, x_{3i}) jointly resampled with replacement $B = 999$ times. From Table 11.1, the 999 bootstrap replication estimates $\widehat{\beta}^*_{3,b}$, $b = 1, \ldots, 999$, had mean 4.716 and standard deviation of 1.939. Table 11.1 also gives key percentiles for $\widehat{\beta}^*_3$ and t^*_3 (defined in the following).

A parametric bootstrap could have been used instead. Then bootstrap samples would be obtained by drawing y_i from the exponential distribution with parameter $\exp(\widehat{\beta}_1 + \widehat{\beta}_2 x_{2i} + \widehat{\beta}_3 x_{3i})$. In the case of tests of $H_0 : \beta_3 = 0$ the exponential parameter could instead be $\exp(\widetilde{\beta}_1 + \widetilde{\beta}_2 x_{2i})$, where $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$ are then the restricted ML estimates from the original sample.

Standard errors: From (11.3) the bootstrap estimate of standard error is computed using the usual standard deviation formula for the 999 bootstrap replication estimates of β_3. This yields estimate 1.939 compared to the usual asymptotic standard error estimate of 1.741. Note that this bootstrap offers no refinement and would only be used as a check or if finding the standard error by other means proved difficult.

Hypothesis testing with asymptotic refinement: We consider test of $H_0 : \beta_3 = 0$ against $H_a : \beta_3 \neq 0$ at level 0.05. A test with asymptotic refinement is based on the t-statistic, which is asymptotically pivotal. From Section 11.2.6 for each bootstrap we compute $t^*_3 = (\widehat{\beta}^*_3 - 4.664)/s_{\widehat{\beta}^*_3}$, which is centered on the estimate $\widehat{\beta}_3 = 4.664$ from the original sample. For a nonsymmetrical test the bootstrap critical values

Table 11.1. *Bootstrap Statistical Inference on a Slope Coefficient: Example*[a]

	$\widehat{\beta}_3^*$	t_3^*	$z = t(\infty)$	$t(47)$
Mean	4.716	0.026	1.021	1.000
SD[b]	1.939	1.047	1.000	1.021
1%	−.336	−2.664	−2.326	−2.408
2.5%	0.501	−2.183	−1.960	−2.012
5%	1.545	−1.728	−1.645	−1.678
25%	3.570	−0.621	−0.675	−0.680
50%	4.772	0.062	0.000	0.000
75%	5.971	0.703	0.675	0.680
95%	7.811	1.706	1.645	1.678
97.5%	8.484	2.066	1.960	2.012
99.0%	9.427	2.529	2.326	2.408

[a] Summary statistics and percentiles based on 999 paired bootstrap resamples for (1) estimate $\widehat{\beta}_3^*$; (2) the associated statistics $t_3^* = (\widehat{\beta}_3^* - \widehat{\beta}_3)/s_{\widehat{\beta}_3^*}$; (3) student t-distribution with 47 degrees of freedom; (4) standard normal distribution. Original dgp is one draw from the exponential distribution given in the text; the sample size is 50.

[b] SD, standard deviation.

equal the lower and upper 2.5 percentiles of the 999 values of t_3^*, the 25th lowest and 25th highest values. From Table 11.1 these are -2.183 and 2.066. Since the t-statistic computed from the original sample $t_3 = (4.664 - 0)/1.741 = 2.679 > 2.066$, the null hypothesis is rejected. A symmetrical test that instead uses the upper 5 percentile of $|t_3^*|$ yields bootstrap critical value 2.078 that again leads to rejection of H_0 at level 0.05.

The bootstrap critical values in this example exceed those using the asymptotic approximation of either standard normal or $t(47)$, an ad hoc finite-sample adjustment motivated by the exact result for linear regression under normality. So the usual asymptotic results in this example lead to overrejection and have actual size that exceeds the nominal size. For example, at 5% the z critical region values of $(-1.960, 1.960)$ are smaller than the bootstrap critical values $(-2.183, 2.066)$. Figure 11.1 plots the bootstrap estimate based on t_3^* of the density of the t-test, smoothed using kernel methods, and compares it to the standard normal. The two densities appear close, though the left tail is notably fatter for the bootstrap estimate. Table 11.1 makes clearer the difference in the tails.

Hypothesis testing without asymptotic refinement: Alternative bootstrap testing methods can be used but do not offer an asymptotic refinement. First, using the bootstrap standard error estimate of 1.939, rather than the asymptotic standard error estimate of 1.741, yields $t_3 = (4.664 - 0)/1.939 = 2.405$. This leads to rejection at level 0.05 using either standard normal or $t(47)$ critical values. Second, from Table 11.1, 95% of the bootstrap estimates $\widehat{\beta}_3^*$ lie in the range (0.501, 8.484), which does not include the hypothesized value of 0, so again we reject $H_0 : \beta_3 = 0$.

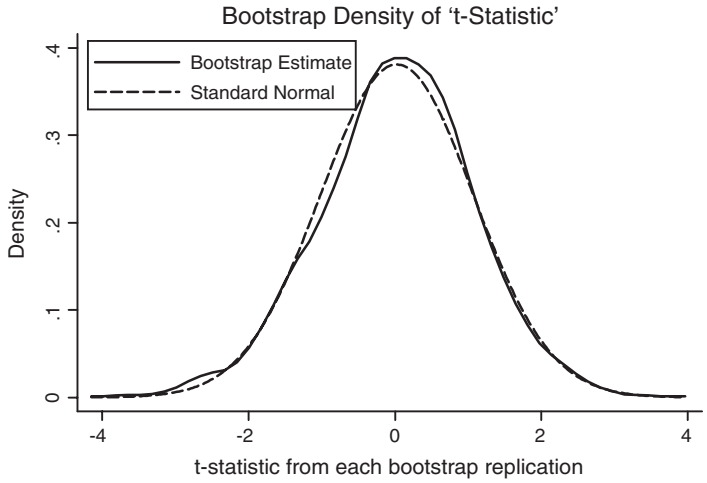

Figure 11.1: Bootstrap density of t-test statistic for slope equal to zero obtained from 999 bootstrap replications with standard normal density plotted for comparison. Data are generated from an exponential distribution regression model.

Confidence intervals: An asymptotic refinement is obtained using the 95% percentile-t confidence interval. Applying (11.6) yields $(4.664 - 2.183 \times 1.741, 4.664 + 2.066 \times 1.741)$ or (0.864, 8.260). This compares to a conventional 95% asymptotic confidence interval of $4.664 \pm 1.960 \times 1.741$ or (1.25, 8.08).

Other confidence intervals can be constructed, but these do not have an asymptotic refinement. Using the bootstrap standard error estimate leads to a 95% confidence interval $4.664 \pm 1.960 \times 1.939 = (0.864, 8.464)$. The percentile method uses the lower and upper 2.5 percentiles of the 999 bootstrap coefficient estimates, leading to a 95% confidence interval of (0.501, 8.484).

Bias correction: The mean of the 999 bootstrap replication estimates of β_3 is 4.716, compared to the original estimate of 4.664. The estimated bias of $(4.716 - 4.664) = 0.052$ is quite small, especially compared to the standard error of $s_3 = 1.741$. The estimated bias is upward and (11.10) yields a bias-corrected estimate of β_3 equal to $4.664 - 0.052 = 4.612$.

The bootstrap relies on asymptotic theory and may actually provide a finite-sample approximation worse than that of conventional methods. To determine that the bootstrap is really an improvement here we need a full Monte Carlo analysis with, say, 1, 000 samples of size 50 drawn from the exponential dgp, with each of these samples then bootstrapped, say, 999 times.

11.4. Bootstrap Theory

The exposition here follows the comprehensive survey of Horowitz (2001). Key results are consistency of the bootstrap and, if the bootstrap is applied to an asymptotically pivotal statistic, asymptotic refinement.

11.4.1. The Bootstrap

We use $X_1, \ldots, X_N$ as generic notation for the data, where for notational simplicity bold is not used for X_i even though it is usually a vector, such as $(y_i, \mathbf{x}_i)$. The data are assumed to be independent draws from distribution with cdf $F_0(x) = \Pr[X \leq x]$. In the simplest applications F_0 is in a finite-dimensional family, with $F_0 = F_0(x, \boldsymbol{\theta}_0)$.

The statistic being considered is denoted $T_N = T_N(X_1, \ldots, X_N)$. The exact finite-sample distribution of T_N is $G_N = G_N(t, F_0) = \Pr[T_N \leq t]$. The problem is to find a good approximation to G_N.

Conventional asymptotic theory uses the asymptotic distribution of T_N, denoted $G_\infty = G_\infty(t, F_0)$. This may theoretically depend on unknown F_0, in which case we use a consistent estimate of F_0. For example, use $\widehat{F}_0 = F_0(\cdot, \widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$.

The empirical bootstrap takes a quite different approach to approximating $G_N(\cdot, F_0)$. Rather than replace G_N by G_∞, the population cdf F_0 is replaced by a consistent estimator F_N of F_0, such as the empirical distribution of the sample.

$G_N(\cdot, F_N)$ cannot be determined analytically but can be approximated by bootstrapping. One bootstrap resample with replacement yields the statistic $T_N^* = T_N(X_1^*, \ldots, X_N^*)$. Repeating this step B independent times yields replications $T_{N,1}^*, \ldots, T_{N,B}^*$. The empirical cdf of $T_{N,1}^*, \ldots, T_{N,B}^*$ is the bootstrap estimate of the distribution of T, yielding

$$\widehat{G}_N(t, F_N) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}(T_{N,b}^* \leq t), \tag{11.11}$$

where $\mathbf{1}(A)$ equals one if event A occurs and equals zero otherwise. This is just the proportion of the bootstrap resamples for which the realized $T_N^* \leq t$.

The notation is summarized in Table 11.2.

11.4.2. Consistency of the Bootstrap

The bootstrap estimate $\widehat{G}_N(t, F_N)$ clearly converges to $G_N(t, F_N)$ as the number of bootstraps $B \to \infty$. Consistency of the bootstrap estimate $\widehat{G}_N(t, F_N)$ for $G_N(t, F_0)$

Table 11.2. *Bootstrap Theory Notation*

Quantity	Notation
Sample (iid)	$X_1, \ldots, X_N$, where X_i is usually a vector
Population cdf of X	$F_0 = F_0(x, \boldsymbol{\theta}_0) = \Pr[X \leq x]$
Statistic of interest	$T_N = T_N(X_1, \ldots, X_N)$
Finite sample cdf of T_N	$G_N = G_N(t, F_0) = \Pr[T_N \leq t]$
Limit cdf of T_N	$G_\infty = G_\infty(t, F_0)$
Asymptotic cdf of T_N	$\widehat{G}_\infty = G_\infty(t, \widehat{F}_0)$, where $\widehat{F}_0 = F_0(x, \widehat{\boldsymbol{\theta}})$
Bootstrap cdf of T_N	$\widehat{G}_N(t, F_N) = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T_{N,b}^* \leq t)$

therefore requires that

$$G_N(t, F_N) \xrightarrow{p} G_N(t, F_0),$$

uniformly in the statistic t and for all F_0 in the space of permitted cdfs.

Clearly, F_N must be consistent for F_0. Additionally, *smoothness* in the dgp $F_0(x)$ is needed, so that $F_N(x)$ and $F_0(x)$ are close to each other uniformly in the observations x for large N. Moreover, *smoothness* in $G_N(\cdot, F)$, the cdf of the statistic considered as a functional of F, is required so that $G_N(\cdot, F_N)$ is close to $G_N(\cdot, F_0)$ when N is large.

Horowitz (2001, pp. 3166–3168) gives two formal theorems, one general and one for iid data, and provides examples of potential failure of the bootstrap, including estimation of the median and estimation with boundary constraints on parameters.

Subject to consistency of F_N for F_0 and smoothness requirements on F_0 and G_N, the bootstrap leads to consistent estimates and asymptotically valid inference. The bootstrap is consistent in a very wide range of settings.

11.4.3. Edgeworth Expansions

An additional attraction of the bootstrap is that it allows for asymptotic refinement. Singh (1981) provided a proof using Edgeworth expansions, which we now introduce.

Consider the asymptotic behavior of $Z_N = \sum_i X_i/\sqrt{N}$, where for simplicity X_i are standardized scalar random variables that are iid [0, 1]. Then application of a central limit theorem leads to a limit standard normal distribution for Z_N. More precisely, Z_N has cdf

$$G_N(z) = \Pr[Z_N \leq z] = \Phi(z) + O(N^{-1/2}), \tag{11.12}$$

where $\Phi(\cdot)$ is the standard normal cdf. The remainder term is ignored and regular asymptotic theory approximates $G_N(z)$ by $G_\infty(z) = \Phi(z)$.

The CLT leading to (11.12) is formally derived by a simple approximation of the **characteristic function** of Z_N, $\mathrm{E}[e^{isZ_N}]$, where $i = -\sqrt{1}$. A better approximation expands this characteristic function in powers of $N^{-1/2}$. The usual **Edgeworth expansion** adds two additional terms, leading to

$$G_N(z) = \Pr[Z_N \leq z] = \Phi(z) + \frac{g_1(z)}{\sqrt{N}} + \frac{g_2(z)}{N} + O(N^{-3/2}), \tag{11.13}$$

where $g_1(z) = -(z^2 - 1)\phi(z)\kappa_3/6$, $\phi(\cdot)$ denotes the standard normal density, κ_3 is the third cumulant of Z_N, and the lengthy expression for $g_2(\cdot)$ is given in Rothenberg (1984, p. 895) or Amemiya (1985, p. 93). In general the rth **cumulant** κ_r is the rth coefficient in the series expansion $\ln(\mathrm{E}[e^{isZ_N}]) = \sum_{r=0}^{\infty} \kappa_r (is)^r/r!$ of the log characteristic function or cumulant generating function.

The remainder term in (11.13) is ignored and an Edgeworth expansion approximates $G_N(z, F_0)$ by $G_\infty(z, F_0) = \Phi(z) + N^{-1/2}g_1(z) + N^{-1}g_2(z)$. If Z_N is a test statistic this can be used to compute p-values and critical values. Alternatively, (11.13) can be

inverted to

$$\Pr\left[Z_N + \frac{h_1(z)}{\sqrt{N}} + \frac{h_2(z)}{N} \leq z\right] \simeq \Phi(z), \tag{11.14}$$

for functions $h_1(z)$ and $h_2(z)$ given in Rothenberg (1984, p. 895). The left-hand side gives a modified statistic that will be better approximated by the standard normal than the original statistic Z_N.

The problem in application is that the cumulants of Z_N are needed to evaluate the functions $g_1(z)$ and $g_2(z)$ or $h_1(z)$ and $h_2(z)$. It can be very difficult to obtain analytical expressions for these cumulants (e.g., Sargan, 1980, and Phillips, 1983). The bootstrap provides a numerical method to implement the Edgeworth expansion without the need to calculate cumulants, as shown in the following.

11.4.4. Asymptotic Refinement via Bootstrap

We now return to the more general setting of Section 11.4.1, with the additional assumption that T_N has a limit normal distribution and usual $\sqrt{N}$ asymptotics apply.

Conventional asymptotic methods use the limit cdf $G_\infty(t, F_0)$ as an approximation to the true cdf $G_N(t, F_0)$. For $\sqrt{N}$-consistent asymptotically normal estimators this has an error that in the limit behaves as a multiple of $N^{-1/2}$. We write this as

$$G_N(t, F_0) = G_\infty(t, F_0) + O(N^{-1/2}), \tag{11.15}$$

where in our example $G_\infty(t, F_0) = \Phi(t)$.

A better approximation is possible using an Edgeworth expansion. Then

$$G_N(t, F_0) = G_\infty(t, F_0) + \frac{g_1(t, F_0)}{\sqrt{N}} + \frac{g_2(t, F_0)}{N} + O(N^{-3/2}). \tag{11.16}$$

Unfortunately, as already noted, the functions $g_1(\cdot)$ and $g_2(\cdot)$ on the right-hand side can be difficult to construct.

Now consider the bootstrap estimator $G_N(t, F_N)$. An Edgeworth expansion yields

$$G_N(t, F_N) = G_\infty(t, F_N) + \frac{g_1(t, F_N)}{\sqrt{N}} + \frac{g_2(t, F_N)}{N} + O(N^{-3/2}); \tag{11.17}$$

see Hall (1992) for details. The bootstrap estimator $G_N(t, F_N)$ is used to approximate the finite-sample cdf $G_N(t, F_0)$. Subtracting (11.16) from (11.17), we get

$$\begin{aligned} G_N(t, F_N) - G_N(t, F_0) = {} & [G_\infty(t, F_N) - G_\infty(t, F_0)] \\ & + \frac{[g_1(t, F_N) - g_1(t, F_0)]}{\sqrt{N}} + O(N^{-1}). \end{aligned} \tag{11.18}$$

Assume that F_N is $\sqrt{N}$ consistent for the true cdf F_0, so that $F_N - F_0 = O(N^{-1/2})$. For continuous function G_∞ the first term on the right-hand side of (11.18), $[G_\infty(t, F_N) - G_\infty(t, F_0)]$, is therefore $O(N^{-1/2})$, so $G_N(t, F_N) - G_N(t, F_0) = O(N^{-1/2})$.

The bootstrap approximation $G_N(t, F_N)$ is therefore in general no closer asymptotically to $G_N(t, F_0)$ than is the usual asymptotic approximation $G_\infty(t, F_0)$; see (11.15).

Now suppose the statistic T_N is *asymptotically pivotal*, so that its asymptotic distribution G_∞ does not depend on unknown parameters. Here this is the case if T_N is standardized so that its limit distribution is the standard normal. Then $G_\infty(t, F_N) = G_\infty(t, F_0)$, so (11.18) simplifies to

$$G_N(t, F_N) - G_N(t, F_0) = N^{-1/2}[g_1(t, F_N) - g_1(t, F_0)] + O(N^{-1}). \tag{11.19}$$

However, because $F_N - F_0 = O(N^{-1/2})$ we have that $[g_1(t, F_N) - g_1(t, F_0)] = O(N^{-1/2})$ for g_1 continuous in F. It follows upon simplification that $G_N(t, F_N) = G_N(t, F_0) + O(N^{-1})$. The bootstrap approximation $G_N(t, F_N)$ is now a better asymptotic approximation to $G_N(t, F_0)$ as the error is now $O(N^{-1})$.

In summary, for a bootstrap on an asymptotically pivotal statistic we have

$$G_N(t, F_0) = G_N(t, F_N) + O(N^{-1}), \tag{11.20}$$

an improvement on the conventional approximation $G_N(t, F_0) = G_\infty(t, F_0) + O(N^{-1/2})$.

The bootstrap on an asymptotically pivotal statistic therefore leads to an improved small-sample performance in the following sense. Let α be the nominal size for a test procedure. Usual asymptotic theory produces t-tests with actual size $\alpha + O(N^{-1/2})$, whereas the bootstrap produces t-tests with actual size $\alpha + O(N^{-1})$.

For symmetric two-sided hypothesis tests and confidence intervals the bootstrap on an asymptotically pivotal statistic can be shown to have approximation error $O(N^{-3/2})$ compared to error $O(N^{-1})$ using usual asymptotic theory.

The preceding results are restricted to asymptotically normal statistics. For chi-squared distributed test statistics the asymptotic gains are similar to those for symmetric two-sided hypothesis tests. For proof of bias reduction by bootstrapping, see Horowitz (2001, p. 3172).

The theoretical analysis leads to the following points. The bootstrap should be from distribution F_N consistent for F_0. The bootstrap requires smoothness and continuity in F_0 and G_N, so that a modification of the standard bootstrap is needed if, for example, there is a discontinuity because of a boundary constraint on the parameters such as $\theta \geq 0$. The bootstrap assumes existence of low-order moments, as low-order cumulants appear in the function g_1 in the Edgeworth expansions. Asymptotic refinement requires use of an asymptotically pivotal statistic. The bootstrap refinement presented assumes iid data, so that modification is needed even for heteroskedastic errors. For more complete discussion see Horowitz (2001).

11.4.5. Power of Bootstrapped Tests

The analysis of the bootstrap has focused on getting tests with correct size in small samples. The size correction of the bootstrap will lead to changes in the power of tests, as will any size correction.

Intuitively, if the actual size of a test using first-order asymptotics exceeds the nominal size, then bootstrapping with asymptotic refinement will not only reduce the size toward the nominal size but, because of less frequent rejection, will also reduce the power of the test. Conversely, if the actual size is less than the nominal size then

bootstrapping will increase test power. This is observed in the simulation exercise of Horowitz (1994, p. 409). Interestingly, in his simulation he finds that although bootstrapping first-order asymptotically equivalent tests leads to tests with similar actual size (essentially equal to the nominal size) there can be considerable difference in test power across the bootstrapped tests.

11.5. Bootstrap Extensions

The bootstrap methods presented so far emphasize smooth $\sqrt{N}$-consistent asymptotically normal estimators based on iid data. The following extensions of the bootstrap permit for a wider range of applications a consistent bootstrap (Sections 11.5.1 and 11.5.2) or a consistent bootstrap with asymptotic refinement (Sections 11.5.3–11.5.5). The presentation of these more advanced methods is brief. Some are used in Section 11.6.

11.5.1. Subsampling Method

The **subsampling method** uses a sample of size m that is substantially smaller than the sample size N. The subsampling may be with replacement (Bickel, Gotze, and van Zwet, 1997) or without replacement (Politis and Romano, 1994).

Replacement subsampling provides subsamples that are random samples of the population, rather than random samples of an estimate of the distribution such as the sample in the case of a paired bootstrap. Replacement subsampling can then be consistent when failure of the smoothness conditions discussed in Section 11.4.2 leads to inconsistency of a full sample bootstrap. The associated asymptotic error for testing or confidence intervals, however, is of higher order of magnitude than the usual $0(N^{-1/2})$ obtained when a full sample bootstrap without refinement can be used.

Subsample bootstraps are useful when full sample bootstraps are invalid, or as a way to verify that a full sample bootstrap is valid. Results will differ with the choice of subsample size. And there is a considerable increase in sample error because a smaller fraction of the sample is being used. Indeed, we should have $(m/N) \to 0$ and $N \to \infty$. Politis, Romano, and Wolf (1999) and Horowitz (2001) provide further details.

11.5.2. Moving Blocks Bootstrap

The **moving blocks bootstrap** is used for data that are dependent rather than independent. This splits the sample into r nonoverlapping blocks of length l, where $rl \simeq N$. First, one samples with replacement from these blocks, to give r new blocks, which will have a different temporal ordering from the original r blocks. Then one estimates the parameters using this bootstrap sample.

The moving blocks method treats the randomly drawn blocks as being independent of each other, but allows dependence within the blocks. A similar blocking was actually used by Anderson (1971) to derive a central limit theorem for an m-dependent process. The moving blocks process requires $r \to \infty$ as $N \to \infty$ to ensure that we

are likely to draw consecutive blocks uncorrelated with each other. It also requires the block length $l \to \infty$ as $N \to \infty$. See, for example, Götze and Künsch (1996).

11.5.3. Nested Bootstrap

A **nested bootstrap**, introduced by Hall (1986), Beran (1987), and Loh (1987), is a bootstrap within a bootstrap. This method is especially useful if the bootstrap is on a statistic that is not asympotically pivotal. In particular, if the standard error of the estimate is difficult to compute one can bootstrap the current bootstrap sample to obtain a bootstrap standard error estimate $s_{\widehat{\theta}^*,\text{Boot}}$ and form $t^* = (\widehat{\theta}^* - \widehat{\theta})/s_{\widehat{\theta}^*,\text{Boot}}$, and then apply the percentile-t method to the bootstrap replications $t_1^*, \ldots, t_B^*$. This permits asymptotic refinements where a single round of bootstrap would not.

More generally, **iterated bootstrapping** is a way to improve the performance of the bootstrap by estimating the errors (i.e., bias) that arise from a single pass of the bootstrap, and correcting for these errors. In general each further iteration of the bootstrap reduces bias by a factor N^{-1} if the statistic is asymptotically pivotal and by a factor $N^{-1/2}$ otherwise. For a good exposition see Hall and Martin (1988). If B bootstraps are performed at each iteration then B^k bootstraps need to be performed if there are k iterations. For this reason at most two iterations, called a **double bootstrap** or **calibrated bootstrap**, are done.

Davison, Hinkley, and Schechtman (1986) proposed **balanced bootstrapping**. This method ensures that each sample observation is reused exactly the same number of times over all B bootstraps, leading to better bootstrap estimates. For implementation see Gleason (1988), whose algorithms add little to computational time compared to the usual unbalanced bootstrap.

11.5.4. Recentering and Rescaling

To yield an asymptotic refinement the bootstrap should be based on an estimate $\widehat{F}$ of the dgp F_0 that imposes all the conditions of the model under consideration. A leading example arises with the residual bootstrap.

Least-squares residuals do not sum to zero in nonlinear models, or even in linear models if there is no intercept. The residual bootstrap (see Section 11.2.4) based on least-squares residuals will then fail to impose the restriction that $\mathrm{E}[u_i] = 0$. The residual bootstrap should instead bootstrap the **recentered residual** $\widehat{u}_i - \bar{u}$, where $\bar{u} = N^{-1} \sum_{i=1}^{N} \widehat{u}_i$. Similar recentering should be done for paired bootstraps of GMM estimators in overidentified models (see Section 11.6.4).

Rescaling of residuals can also be useful. For example, in the linear regression model with iid errors resample from $(N/(N-K))^{1/2}\widehat{u}_i$ since these have variance s^2. Other adjustments include using the standardized residual $\widehat{u}_i/\sqrt{(1-h_{ii})s^2}$, where h_{ii} is the ith diagonal entry in the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

11.5.5. The Jackknife

The bootstrap can be used for bias correction (see Section 11.2.8). An alternative resampling method is the jackknife, a precursor of the bootstrap. The jackknife uses N

deterministically defined subsamples of size $N-1$ obtained by dropping in turn each of the N observations and recomputing the estimator.

To see how the jackknife works, let $\widehat{\theta}_N$ denote the estimate of θ using all N observations, and let $\widehat{\theta}_{N-1}$ denote the estimate of θ using the first $(N-1)$ observations. If (11.7) holds then $\mathrm{E}[\widehat{\theta}_N] = \theta + a_N/N + b_N/N^2 + O(N^{-3})$ and $\mathrm{E}[\widehat{\theta}_{N-1}] = \theta + a_N/(N-1) + b_N/(N-1)^2 + O(N^{-3})$, which implies $\mathrm{E}[N\widehat{\theta}_N - (N-1)\widehat{\theta}_{N-1}] = \theta + O(N^{-2})$. Thus $N\widehat{\theta}_N - (N-1)\widehat{\theta}_{N-1}$ has smaller bias than $\widehat{\theta}_N$.

The estimator can be more variable, however, as it uses less of the data. As an extreme example, if $\widehat{\theta} = \bar{y}$ then the new estimator is simply y_N, the Nth observation. The variation can be reduced by dropping each observation in turn and averaging.

More formally then, consider the estimator $\widehat{\boldsymbol{\theta}}$ of a parameter vector $\boldsymbol{\theta}$ based on a sample of size N from iid data. For $i = 1, \ldots, N$ sequentially delete the ith observation and obtain N jackknife replication estimates $\widehat{\boldsymbol{\theta}}_{(-i)}$ from the N jackknife resamples of size $(N-1)$. The **jackknife estimate of the bias** of $\widehat{\boldsymbol{\theta}}$ is $(N-1)(\overline{\widehat{\boldsymbol{\theta}}} - \widehat{\boldsymbol{\theta}})$, where $\overline{\widehat{\boldsymbol{\theta}}} = N^{-1}\sum_i \widehat{\boldsymbol{\theta}}_{(-i)}$ is the average of the N jackknife replications $\widehat{\boldsymbol{\theta}}_{(-i)}$. The bias appears large because of multiplication by $(N-1)$, but the differences $(\widehat{\boldsymbol{\theta}}_{(-i)} - \widehat{\boldsymbol{\theta}})$ are much smaller than in the bootstrap case since a jackknife resample differs from the original sample in only one observation.

This leads to the bias-corrected **jackknife estimate** of $\boldsymbol{\theta}$:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\text{Jack}} &= \widehat{\boldsymbol{\theta}} - (N-1)(\overline{\widehat{\boldsymbol{\theta}}} - \widehat{\boldsymbol{\theta}}) \\ &= N\widehat{\boldsymbol{\theta}} - (N-1)\overline{\widehat{\boldsymbol{\theta}}}. \end{aligned} \tag{11.21}$$

This reduces the bias from $O(N^{-1})$ to $O(N^{-2})$, which is the same order of bias reduction as for the bootstrap. It is assumed that, as for the bootstrap, the estimator is a smooth $\sqrt{N}$-consistent estimator. The jackknife estimate can have increased variance compared with $\widehat{\boldsymbol{\theta}}$, and examples where the jackknife fails are given in Miller (1974).

A simple example is estimation of σ^2 from an iid sample with $y_i \sim [\mu, \sigma^2]$. The estimate $\widehat{\sigma}^2 = N^{-1}\sum_i (y_i - \bar{y})^2$, the MLE under normality, has $\mathrm{E}[\widehat{\sigma}^2] = \sigma^2(N-1)/N$ so that the bias equals σ^2/N, which is $O(N^{-1})$. In this example the jackknife estimate can be shown to simplify to $\widehat{\sigma}^2_{\text{Jack}} = (N-1)^{-1}\sum_i (y_i - \bar{y})^2$, so one does not need not to compute N separate estimates $\widehat{\sigma}^2_{(-i)}$. This is an unbiased estimate of σ^2, so the bias is actually zero rather than the general result of $O(N^{-2})$.

The jackknife is due to Quenouille (1956). Tukey (1958) considered application to a wider range of statistics. In particular, the **jackknife estimate of the standard error** of an estimator $\widehat{\theta}$ is

$$\widehat{\text{se}}_{\text{Jack}}[\widehat{\theta}] = \left[\frac{N-1}{N} \sum_{i=1}^{N} (\widehat{\theta}_{(-i)} - \overline{\widehat{\theta}})^2 \right]^{1/2}. \tag{11.22}$$

Tukey proposed the term jackknife by analogy to a Boy Scout jackknife that solves a variety of problems, each of which could be solved more efficiently by a specially constructed tool. The jackknife is a "rough and ready" method for bias reduction in many situations, but it is not the ideal method in any. The jackknife can be viewed as a linear approximation of the bootstrap (Efron and Tibsharani, 1993, p. 146). It requires

less computation than the bootstrap in small samples, as then $N < B$ is likely, but is outperformed by the bootstrap as $B \to \infty$.

Consider the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, with $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$. An example of a biased estimator from OLS regression is a time-series model with lagged dependent variable as regressor. The regression estimator based on the ith jackknife sample $(\mathbf{X}_{(-i)}, \mathbf{y}_{(-i)})$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(-i)} &= [\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)} \\ &= [\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i]^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{x}_i y_i) \\ &= \widehat{\boldsymbol{\beta}} - [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}_i(y_i - \mathbf{x}'_i\widehat{\boldsymbol{\beta}}_{(-i)}).\end{aligned}$$

The third equality avoids the need to invert $\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}$ for each i and is obtained using

$$[\mathbf{X}'\mathbf{X}]^{-1} = [\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1} - \frac{[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}]^{-1}\mathbf{x}_i\mathbf{x}'_i\left[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}\right]^{-1}}{1 + \mathbf{x}'_i\left[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}\right]^{-1}\mathbf{x}_i}.$$

Here the **pseudo-values** are given by $N\widehat{\boldsymbol{\beta}} - (N-1)\,\widehat{\boldsymbol{\beta}}_{(-i)}$, and the jackknife estimator of $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\text{Jack}} = N\widehat{\boldsymbol{\beta}} - (N-1)\frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\beta}}_{(-i)}. \tag{11.23}$$

An interesting application of the jackknife to bias reduction is the jackknife IV estimator (see Section 6.4.4).

11.6. Bootstrap Applications

We consider application of the bootstrap taking into account typical microeconometric complications such as heteroskedasticity and clustering and more complicated estimators that can lead to failure of simple bootstraps.

11.6.1. Heteroskedastic Errors

For least squares in models with additive errors that are heteroskedastic, the standard procedure is to use White's heteroskedastic-consistent covariance matrix estimator (HCCME). This is well known to perform poorly in small samples. When done correctly, the bootstrap can provide an improvement.

The paired bootstrap leads to valid inference, since the essential assumption that $(y_i, \mathbf{x}_i)$ is iid still permits $\mathrm{V}[u_i|\mathbf{x}_i]$ to vary with $\mathbf{x}_i$ (see Section 4.4.7). However, it does not offer an asymptotic refinement because it does not impose the condition that $\mathrm{E}[u_i|\mathbf{x}_i] = 0$.

The usual residual bootstrap actually leads to invalid inference, since it assumes that $u_i|\mathbf{x}_i$ is iid and hence erroneously imposes the condition of homoskedastic errors. In terms of Section 11.4 theory, $\widehat{F}$ is then inconsistent for F. One can specify a formal model for heteroskedasticity, say $u_i = \exp(\mathbf{z}'_i\alpha)\varepsilon_i$, where ε_i are iid, obtain estimate $\exp(\mathbf{z}'_i\widehat{\alpha})$, and then bootstrap the implied residuals $\widehat{\varepsilon}_i$. Consistency and asymptotic

refinement of this bootstrap requires correct specification of the functional form for the heteroskedasticity.

The **wild bootstrap**, introduced by Wu (1986) and Liu (1988) and studied further by Mammen (1993), provides asymptotic refinement without imposing such structure on the heteroskedasticity. This bootstrap replaces the OLS residual $\widehat{u}_i$ by the following residual:

$$\widehat{u}_i^* = \begin{cases} \frac{1-\sqrt{5}}{2}\widehat{u}_i \simeq -0.6180\widehat{u}_i & \text{with probability } \frac{1+\sqrt{5}}{2\sqrt{5}} \simeq 0.7236, \\ [1 - \frac{1-\sqrt{5}}{2}]\widehat{u}_i \simeq 1.6180\widehat{u}_i & \text{with probability } 1 - \frac{1+\sqrt{5}}{2\sqrt{5}} \simeq 0.2764. \end{cases}$$

Taking expectations with respect to only this two-point distribution and performing some algebra yields $\mathrm{E}[\widehat{u}_i^*] = 0$, $\mathrm{E}[\widehat{u}_i^{*2}] = \widehat{u}_i^2$, and $\mathrm{E}[\widehat{u}_i^{*3}] = \widehat{u}_i^3$. Thus $\widehat{u}_i^*$ leads to a residual with zero conditional mean as desired, since $\mathrm{E}[\widehat{u}_i^*|\widehat{u}_i, \mathbf{x}_i] = 0$ implies $\mathrm{E}[\widehat{u}_i^*|\mathbf{x}_i] = 0$, while the second and third moments are unchanged.

The wild bootstrap resamples have ith observation $(y_i^*, \mathbf{x}_i)$, where $y_i^* = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{u}_i^*$. The resamples vary because of different realizations of $\widehat{u}_i^*$. Simulations by Horowitz (1997, 2001) show that this bootstrap works much better than a paired bootstrap when there is heteroskedasticity and works well compared to other bootstrap methods even if there is no heteroskedasticity.

It seems surprising that this bootstrap should work because for the ith observation it draws from only two possible values for the residual, $-0.6180\widehat{u}_i$ or $1.6180\widehat{u}_i$. However, a similar draw is being made over all N observations and over B bootstrap iterations. Recall also that White's estimator replaces $\mathrm{E}[u_i^2]$ by $\widehat{u}_i^2$, which, although incorrect for one observation, is valid when averaged over the sample. The wild bootstrap is instead drawing from a two-point distribution with mean 0 and variance $\widehat{u}_i^2$.

11.6.2. Panel Data and Clustered Data

Consider a linear panel regression model

$$\widetilde{y}_{it} = \widetilde{\mathbf{w}}_{it}'\boldsymbol{\theta} + \widetilde{u}_{it},$$

where i denotes individual and t denotes time period. Following the notation of Section 21.2.3, the tilda is added as the original data y_{it} and $\mathbf{x}_{it}$ may first be transformed to eliminate fixed effects, for example. We assume that the errors $\widetilde{u}_{it}$ are independent over i, though they may be heteroskedastic and correlated over t for given i.

If the panel is short, so that T is finite and asymptotic theory relies on $N \to \infty$, then consistent standard errors for $\widehat{\boldsymbol{\theta}}$ can be obtained by a paired or EDF bootstrap that resamples over i but does not resample over t. In the preceding presentation $\mathbf{w}_i$ becomes $[y_{i1}, \mathbf{x}_{i1}, \ldots, y_{iT}, \mathbf{x}_{iT}]$ and we resample over i and obtain all T observations for the chosen i.

This **panel bootstrap**, also called a block bootstrap, can also be applied to the nonlinear panel models of Chapter 23. The key assumptions are that the panel is short and the data are independent over i. More generally, this bootstrap can be applied whenever data are clustered (see Section 24.5), provided cluster size is finite and the number of clusters goes to infinity.

The panel bootstrap produces standard errors that are asymptotically equivalent to panel robust sandwich standard errors (see Section 21.2.3). It does not provide an asymptotic refinement. However, it is quite simple to implement and is practically very useful as many packages do not automatically provide panel robust standard errors even for quite basic panel estimators such as the fixed effects estimator. Depending on the application, other bootstraps such as parametric and residual bootstraps may be possible, provided again that resampling is over i only.

Asymptotic refinement is straightforward if the errors are iid. More realistically, however, $\widetilde{u}_{it}$ will be heteroskedastic and correlated over t for given i. The wild bootstrap (see Section 11.6.1) should provide an asymptotic refinement in a linear model if the panel is short. Then wild bootstrap resamples have (i, t)th observation $(\widetilde{y}^*_{it}, \widetilde{\mathbf{w}}_{it})$, where $\widetilde{y}^*_{it} = \widetilde{\mathbf{w}}_{it}'\boldsymbol{\theta} + \widehat{\widetilde{u}}^*_{it}$, $\widehat{\widetilde{u}}_{it} = \widetilde{y}_{it} - \widetilde{\mathbf{w}}_{it}'\widehat{\boldsymbol{\theta}}$ and $\widehat{\widetilde{u}}^*_{it}$ is a draw from the two-point distribution given in Section 11.6.1.

11.6.3. Hypothesis and Specification Tests

Section 11.2.6 focused on tests of the hypothesis $\theta = \theta_0$. Here we consider more general tests. As in Section 11.2.6, the bootstrap can be used to perform hypothesis tests with or without asymptotic refinement.

Tests without Asymptotic Refinement

A leading example of the usefulness of the bootstrap is the Hausman test (see Section 8.3). Standard implementation of this test requires estimation of $\mathrm{V}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}]$, where $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ are the two estimators being contrasted. Obtaining this estimate can be difficult unless the strong assumption is made that one of the estimators is fully efficient under H_0. The paired bootstrap can be used instead, leading to consistent estimate

$$\widehat{\mathrm{V}}_{\mathrm{Boot}}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}] = \frac{1}{B-1}\sum_{b=1}^{B}[(\widehat{\boldsymbol{\theta}}^*_b - \widetilde{\boldsymbol{\theta}}^*_b) - (\overline{\widehat{\boldsymbol{\theta}}}^* - \overline{\widetilde{\boldsymbol{\theta}}}^*)][(\widehat{\boldsymbol{\theta}}^*_b - \widetilde{\boldsymbol{\theta}}^*_b) - (\overline{\widehat{\boldsymbol{\theta}}}^* - \overline{\widetilde{\boldsymbol{\theta}}}^*)]',$$

where $\overline{\widehat{\boldsymbol{\theta}}}^* = B^{-1}\sum_b \widehat{\boldsymbol{\theta}}^*_b$ and $\overline{\widetilde{\boldsymbol{\theta}}}^* = B^{-1}\sum_b \widetilde{\boldsymbol{\theta}}^*_b$. Then compute

$$\mathrm{H} = (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'\left(\widehat{\mathrm{V}}_{\mathrm{Boot}}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}]\right)^{-1}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) \tag{11.24}$$

and compare to chi-square critical values. As mentioned in Chapter 8, a generalized inverse may need to be used and care may be needed to ensure chi-square critical values are obtained using the correct degrees of freedom.

More generally, this approach can be used for any standard normal test or chi-square distributed test where implementation is difficult because a variance matrix must be estimated. Examples include hypothesis tests based on a two-step estimator and the m-tests of Chapter 8.

Tests with Asymptotic Refinement

Many tests, especially those for fully parametric models such as the LM test and IM test, can be simply implemented using an auxiliary regression (see Sections 7.3.5 and

8.2.2). The resulting test statistics, however, perform poorly in finite samples as documented in many Monte Carlo studies. Such test statistics are easily computed and are asymptotically pivotal as the chi-square distribution does not depend on unknown parameters. They are therefore prime candidates for asymptotic refinement by bootstrap.

Consider the m-test of $H_0 : \mathrm{E}[\mathbf{m}_i(y_i|\mathbf{x}_i, \boldsymbol{\theta})] = \mathbf{0}$ against $H_a : \mathrm{E}[\mathbf{m}_i(y_i|\mathbf{x}_i, \boldsymbol{\theta})] \neq \mathbf{0}$ (see Section 8.2). From the original data estimate $\widehat{\boldsymbol{\theta}}$ by ML, and calculate the test statistic M. Using a parametric bootstrap, resample y_i^* from the fitted conditional density $f(y_i|\mathbf{x}_i, \widehat{\boldsymbol{\theta}})$, for fixed regressors in repeated samples, or from $f(y_i|\mathbf{x}_i^*, \widehat{\boldsymbol{\theta}})$. Compute M_b^*, $b = 1, \ldots, B$, in the bootstrap resamples. Reject H_0 at level α if the original calculated statistic M exceeds the α quantile of M_b^*, $b = 1, \ldots, B$.

Horowitz (1994) presented this bootstrap for the IM test and demonstrated with simulation examples that there are substantial finite-sample gains to this bootstrap. A detailed application by Drukker (2002) to specification tests for the tobit model suggests that conditional moment specification tests can be easily applied to fully parametric models, since any size distortion in the auxiliary regressions can be corrected through bootstrap.

Note that bootstrap tests without asymptotic refinement, such as the Hausman test given here, can be refined by use of the nested bootstrap given in Section 11.5.3.

11.6.4. GMM, Minimum Distance, and Empirical Likelihood in Overidentified Models

The GMM estimator is based on population moment conditions $\mathrm{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$ (see Section 6.3.1). In a just-identified model a consistent estimator simply solves $N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}$. In overidentified models this estimator is no longer feasible. Instead, the GMM estimator is used (see Section 6.3.2).

Now consider bootstrapping, using the paired or EDF bootstrap. For GMM in an *overidentified* model $N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) \neq \mathbf{0}$, so this bootstrap does not impose on the bootstrap resamples the original population restriction that $\mathrm{E}[\mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$. As a result even if the asymptotically pivotal t-statistic is used there is no longer a bootstrap refinement, though bootstraps on $\widehat{\boldsymbol{\theta}}$ and related confidence intervals and t-test statistics remain consistent. More fundamentally, the bootstrap of the OIR test (see Section 6.3.8) can be shown to be inconsistent. We focus on cross-section data but similar issues arise for panel GMM estimators (see Chapter 22) in overidentified models.

Hall and Horowitz (1996) propose correcting this by **recentering**. Then the bootstrap is based on $\mathbf{h}^*(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}) - N^{-1} \sum_i \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$ and asymptotic refinements can be obtained for statistics based on $\widehat{\boldsymbol{\theta}}$ including the OIR test.

Horowitz (1998) does similar recentering for the minimum distance estimator (see Section 6.7). He then applies the bootstrap to the covariance structure example of Altonji and Segal (1996) discussed in Section 6.3.5.

An alternative adjustment proposed by Brown and Newey (2002) is to not recenter but to instead resample the observations $\mathbf{w}_i$ with probabilities that vary across observations rather than using equal weights $1/N$. Specifically, let $\Pr[\mathbf{w}^* = \mathbf{w}_i] = \widehat{\pi}_i$, where $\widehat{\pi}_i = (1 + \widehat{\boldsymbol{\lambda}}' \widehat{\mathbf{h}}_i)$, $\widehat{\mathbf{h}}_i = \mathbf{h}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$, and $\widehat{\boldsymbol{\lambda}}$ maximizes $\sum_i \ln(1 + \widehat{\boldsymbol{\lambda}}' \widehat{\mathbf{h}}_i)$. The motivation is that the probabilities $\widehat{\pi}_i$ equivalently are the solution to an empirical likelihood (EL)

problem (see Section 6.8.2) of maximizing $\sum_i \ln \pi_i$ with respect to $\pi_1, \ldots, \pi_N$ subject to the constraints $\sum_i \pi_i \widehat{\mathbf{h}}_i = \mathbf{0}$ and $\sum_i \pi_i = 1$. This **empirical likelihood bootstrap** of the GMM estimator therefore imposes the constraint $\sum_i \widehat{\pi}_i \widehat{\mathbf{h}}_i = \mathbf{0}$.

One could instead work directly with EL from the beginning, letting $\widehat{\boldsymbol{\theta}}$ be the EL estimator rather than the GMM estimator. The advantage of the Brown and Newey (2002) approach is that it avoids the more challenging computation of the EL estimator. Instead, one needs only the GMM estimator and solution of the concave programming problem of minimizing $\sum_i \ln(1 + \widehat{\boldsymbol{\lambda}}' \widehat{\mathbf{h}}_i)$.

11.6.5. Nonparametric Regression

Nonparametric density and regression estimators converge at rate less than $\sqrt{N}$ and are asymptotically biased. This complicates inference such as confidence intervals (see Sections 9.3.7 and 9.5.4).

We consider the kernel regression estimator $\widehat{m}(x_0)$ of $m(x_0) = \mathrm{E}[y|x = x_0]$ for observations (y, x) that are iid, though conditional heteroskedasticity is permitted. From Horowitz (2001, p. 3204), an asymptotically pivotal statistic is

$$t = \frac{\widehat{m}(x_0) - m(x_0)}{s_{\widehat{m}(x_0)}},$$

where $\widehat{m}(x_0)$ is an undersmoothed kernel regression estimator with bandwidth $h = o(N^{-1/3})$ rather than the optimal $h^* = O(N^{-1/5})$ and

$$s^2_{\widehat{m}(x_0)} = \frac{1}{Nh[\widehat{f}(x_0)]^2} \sum_{i=1}^{N} (y_i - \widehat{m}(x_i))^2 K\left(\frac{x_i - x_0}{h}\right)^2,$$

where $\widehat{f}(x_0)$ is a kernel estimate of the density $f(x)$ at $x = x_0$. A paired bootstrap resamples (y^*, x^*) and forms $t_b^* = [\widehat{m}_b^*(x_0) - m(x_0)]/s^*_{\widehat{m}(x_0),b}$, where $s^*_{\widehat{m}(x_0),b}$ is computed using bootstrap sample kernel estimates $\widehat{m}_b^*(x_i)$ and $\widehat{f}_b^*(x_0)$. The percentile-t confidence interval of Section 11.2.7 then provides an asymptotic refinement. For a symmetrical confidence interval or symmetrical test at level α the error is $o((Nh^{-1}))$ rather than $O((Nh^{-1}))$ using first-order asymptotic approximation.

Several variations on this bootstrap are possible. Rather than using undersmoothing, bias can be eliminated by directly estimating the bias term given in Section 9.5.2. Also rather than using $s^2_{\widehat{m}(x_0)}$, the variance term given in Section 9.5.2 can be directly estimated.

Yatchew (2003) provides considerable detail on implementing the bootstrap in nonparametric and semiparametric regression.

11.6.6. Nonsmooth Estimators

From Section 11.4.2 the bootstrap assumes smoothness in estimators and statistics. Otherwise the bootstrap may not offer an asymptotic refinement and may even be invalid.

As illustration we consider the LAD estimator and extension to binary data. The LAD estimator (see Section 4.6.2) has objective function $\sum_i |y_i - \mathbf{x}_i'\boldsymbol{\beta}|$ that has

discontinuous first derivative. A bootstrap can provide a valid asymptotic approximation but does not provide an asymptotic refinement. For binary outcomes, the LAD estimator extends to the maximum score estimator of Manski (1975) (see Section 14.7.2). For this estimator the bootstrap is not even consistent.

In these examples bootstraps with asymptotic refinements can be obtained by using a smoothed version of the original objective function for the estimator. For example, the smoothed maximum score estimator of Horowitz (1992) is presented in Section 14.7.2.

11.6.7. Time Series

The bootstrap relies on resampling from an iid distribution. Time-series data therefore present obvious problems as the result of dependence.

The bootstrap is straightforward in the linear model with an ARMA error structure and resampling the underlying white noise error. As an example, suppose $y_t = \beta x_t + u_t$, where $u_t = \rho u_{t-1} + \varepsilon_t$ and ε_t is white noise. Then given estimates $\widehat{\beta}$ and $\widehat{\rho}$ we can recursively compute residuals as $\widehat{\varepsilon}_t = \widehat{u}_t - \widehat{\rho}\widehat{u}_{t-1} = y_t - x_t\widehat{\beta} - \widehat{\rho}(y_{t-1} - x_{t-1}\widehat{\beta})$. Bootstrapping these residuals to give $\widehat{\varepsilon}_t^*$, $t = 1, \ldots, T$, we can then recursively compute $\widehat{u}_t^* = \rho\widehat{u}_{t-1}^* + \widehat{\varepsilon}_t^*$ and hence $y_t^* = \widehat{\beta}x_t + \widehat{u}_t^*$. Then regress y_t^* on x_t with AR(1) error. An early example was presented by Freedman (1984), who bootstrapped a dynamic linear simultaneous equations regression model estimated by 2SLS. Given linearity, simultaneity adds little problems. The dynamic nature of the model is handled by recursively constructing $\mathbf{y}_t^* = f(\mathbf{y}_{t-1}^*, \mathbf{x}_t, \mathbf{u}_t^*)$, where $\mathbf{u}_t^*$ are obtained by resampling from the 2SLS structural equation residuals and $\mathbf{y}_0^* = \mathbf{y}_0$. Then perform 2SLS on each bootstrap sample.

This method assumes the underlying error is iid. For general dependent data without an ARMA specification, for example, nonstationary data, the moving blocks bootstrap presented in Section 11.5.2 can be used.

For testing unit roots or cointegration special care is needed in applying the bootstrap as the behavior of the test statistic changes discontinuously at the unit root. See, for example, Li and Maddala (1997). Although it is possible to implement a valid bootstrap in this situation, to date these bootstraps do not provide an asymptotic refinement.

11.7. Practical Considerations

The bootstrap without asymptotic refinement can be a very useful tool for the applied researcher in situations where it is difficult to perform inference by other means. This need can vary with available software and the practitioner's tool kit. The most common application of the bootstrap to date is computation of standard errors needed to conduct a Wald hypothesis test. Examples include heteroskedasticity-robust and panel-robust inference, inference for two-step estimators, and inference on transformations of estimators. Other potential applications include computation of m-test statistics such as the Hausman test.

The bootstrap can additionally provide an asymptotic refinement. Many Monte Carlo studies show that quite standard procedures can perform poorly in finite samples. There appears to be great potential for use of bootstrap refinements, currently unrealized. In some cases this could improve existing inference, such as use of the wild bootstrap in models with additive errors that are heteroskedastic. In other cases it should encourage increased use of methods that are currently under-utilized. In particular, model specification tests with good small-sample properties can be implemented by bootstrapping easily computed auxiliary regressions.

There are two barriers to the use of the bootstrap. First, the bootstrap is not always built into statistical packages. This will change over time, and for now constructing code for a bootstrap is not too difficult provided the package includes looping and the ability to save regression output. Second, there are subtleties involved. Asymptotic refinement requires use of an asymptotically pivotal statistic and the simplest bootstraps presume iid data and smoothness of estimators and statistics. This covers a wide class of applications but not all applications.

11.8. Bibliographic Notes

The bootstrap was proposed by Efron (1979) for the iid case. Singh (1981) and Bickel and Freedman (1981) provided early theory. A good introductory statistics treatment is by Efron and Tibsharani (1993), and a more advanced treatment is by Hall (1992). Extensions to the regression case were considered early on; see, for example, Freedman (1984). Most of the work by econometricians has occurred in the past 10 years. The survey of Horowitz (2001) is very comprehensive and is well complemented by the survey of Brownstone and Kazimi (1998), which considers many econometrics applications, and the paper by MacKinnon (2002).

Exercises

11–1 Consider the model $y = \alpha + \beta x + \varepsilon$, where α, β, and x are scalars and $\varepsilon \sim \mathcal{N}[0, \sigma^2]$. Generate a sample of size $N = 20$ with $\alpha = 2$, $\beta = 1$, and $\sigma^2 = 1$ and suppose that $x \sim \mathcal{N}[2, 2]$. We wish to test $H_0 : \beta = 1$ against $H_a : \beta \neq 1$ at level 0.05 using the t-statistic $t = (\widehat{\beta} - 1)/\text{se}[\widehat{\beta}]$. Do as much of the following as your software permits. Use $B = 499$ bootstrap replications.

- **(a)** Estimate the model by OLS, giving slope estimate $\widehat{\beta}$.
- **(b)** Use a paired bootstrap to compute the standard error and compare this to the original sample estimate. Use the bootstrap standard error to test H_0.
- **(c)** Use a paired bootstrap with asymptotic refinement to test H_0.
- **(d)** Use a residual bootstrap to compute the standard error and compare this to the original sample estimate. Use the bootstrap standard error to test H_0.
- **(e)** Use a residual bootstrap with asymptotic refinement to test H_0.

11–2 Generate a sample of size 20 according from the following dgp. The two regressors are generated by $x_1 \sim \chi^2(4) - 4$ and $x_2 \sim 3.5 + \mathcal{U}[1, 2]$; the error is from a mixture of normals with $u \sim \mathcal{N}[0, 25]$ with probability 0.3 and $u \sim \mathcal{N}[0, 5]$ with

probability 0.7; and the dependent variable is $y = 1.3x_1 + 0.7x_2 + 0.5u$.

(a) Estimate by OLS the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$.

(b) Suppose we are interested in estimating the quantity $\gamma = \beta_1 + \beta_2^2$ from the data. Use the least-squares estimates to estimate this quantity. Use the delta method to obtain approximate standard error for this function.

(c) Then estimate the standard error of $\widehat{\gamma}$ using a paired bootstrap. Compare this to se$[\widehat{\gamma}]$ from part (b) and explain the difference. For the bootstrap use $B = 25$ and $B = 200$.

(d) Now test $H_0 : \gamma = 1.0$ at level 0.05 using a paired bootstrap with $B = 999$. Perform bootstrap tests without and with asymptotic refinement.

11–3 Use 200 observations from the Section 4.6.4 data on natural logarithm of health expenditure (y) and natural logarithm of total expenditure (x). Obtain OLS estimates of the model $y = \alpha + \beta x + u$. Use the paired bootstrap with $B = 999$.

(a) Obtain a bootstrap estimate of the standard error of $\widehat{\beta}$.

(b) Use this standard error estimate to test $H_0 : \beta = 1$ against $H_a : \beta \neq 1$.

(c) Do a bootstrap test with refinement of $H_0 : \beta = 1$ against $H_a : \beta \neq 1$ under the assumption that u is homoskedastic.

(d) If u is heteroskedastic what happens to your method in (c)? Is the test still asymptotically valid, and if so does it offer an asymptotic refinement?

(e) Do a bootstrap to obtain a bias-corrected estimate of β.