Economics 675 – Fall 2011
Professor Jeffrey Smith and Robert Garlick
Bootstrapping Lecture
Version of November 23, 2011

Motivation

The bootstrap is an alternative to using analytically derived estimators, typically based on first-order asymptotic approximations. It is most commonly used to derive variance estimators or test statistics but may in principle be used in a wide range of other contexts. The bootstrap is particularly attractive when:

1) It is difficult to derive analytical expressions for estimators. The bootstrap avoids the need for this derivation.
   Example: Two-step estimators, where the asymptotic variance should reflect uncertainty in both steps.

2) Analytical expressions are available but are biased in finite samples. Bootstrap-based statistics may be less biased.
   Example: The cluster-robust variance estimator converges with the number of clusters, not observations. It may have large finite sample bias when the number of clusters is small, even if the number of observations is large.

3) Analytical expressions are available but the asymptotic distribution of the estimator is difficult to calculate or is not standard normal. In this case, tests based on critical values from the standard normal distribution may be misleading. The bootstrap allows us approximate the distribution of test statistics directly, rather than relying on the standard normal approximation.

Intuitive overview

Recall that the sampling distribution of a random variable (including a statistic) describes its behavior in repeated random samples from its population distribution (or from the underlying population distribution on which the statistic is based).

We can estimate the sampling distribution of a random variable by examining the behavior of the random variable in repeated random samples drawn from the population distribution. In practice, we seldom have access to repeated random samples. But if an *estimate* of the population distribution is available, we can take repeated random samples from this estimate.

The key idea of the bootstrap is to estimate the population distribution, $F(X)$, with the sample distribution,

$$\hat{F}(X) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq X\}$$

Equivalently, we can estimate the population density/mass function, $f(X)$, with the sample mass function,

$$\hat{f}(X) = \begin{cases} \frac{1}{n} & if\ X = X_i \\ 0 & otherwise \end{cases}$$

This is a simple extension of the *analogue principle*, in which we estimate population parameters (mean, variance, etc.) with their sample analogues.

[Draw two pictures. First, a continuous population density function and the corresponding estimated density (a series of mass points at the values of the random variable in the sample). Second, a discrete mass function and the corresponding estimated density. Note that if this is binomial, the population is fully characterized by a single parameter.]

To obtain a bootstrap estimate of the variance of a statistic (or a random variable) we draw repeated samples from the estimated distribution, density or mass function and compute the variance of the statistic across these samples.

Some notation

Let $X_1,...,X_n$ denote the observed sample, θ denote the parameter of interest, and $\hat{\theta} = s(X)$ denote the realized value of an estimator of the parameter of interest, calculated on the observed data.

Similarly, let $X_1^*,...,X_n^*$ denote a bootstrap sample drawn from $\hat{f}(X)$ and let $\hat{\theta}^* = s(X^*)$ denote the realization of the estimator calculated on the bootstrap sample.

Algorithm

The algorithm for estimating the bootstrap standard errors is then the following:

1. Using the original data, construct an estimate of the parameter of interest, $\hat{\theta}$.
2. For b = 1 to B, where B is the total number of bootstrap samples
    a. Pick a bootstrap sample $X_1^{*b},...,X_n^{*b}$ of size n from $\hat{f}(X)$ with replacement.
    b. Calculate $\hat{\theta}^{*b} = s(X^{*b})$.
3. Estimate the bootstrap standard error as the standard deviation of the parameter estimates from the B bootstrap samples. In notation:

$$\widehat{se}_B(\hat{\theta}) = \left(\frac{\sum_{b=1}^{B} \left[\hat{\theta}^{*b} - \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b} \right]}{B-1} \right)^{1/2}$$

Under appropriate assumptions, the bootstrap standard error is consistent for the square root of the asymptotic variance of $\hat{\theta}$. These assumptions hold in most econometric applications but several important exceptions are discussed below.

Applications

We consider three sets of applications. These are not intended to be exhaustive of all possible applications but these cover some of the most common uses of the bootstrap.

Variance estimation

Some commonly used estimators have asymptotic variances that are difficult to calculate. We typically estimate the q^{th} quantile of a continuous distribution with the consistent estimator $\hat{F}^{-1}(q) = \min\{X_i: \hat{F}(X_i) \leq q\}$. However, this estimator's asymptotic variance is

$$\frac{q(1-q)}{n[f(F^{-1}(q))]^2}$$

which requires a consistent estimator of the density function in order to implement. Even if this is available, the resultant variance estimator will be biased. (Remember that the expectation of a ratio is not equal to the ratio of the expectations!) We may therefore prefer a bootstrap-based estimator of the variance.

Chernozhukov, Fernandez-Val and Melly (2009, mimeo) develop a framework for estimating counterfactual distributions: the hypothetical marginal distribution of outcomes in population A if they had the same distribution of covariates as population B. They then consider summary statistics of interest for the counterfactual distribution (moments, order statistics, inequality indices, etc.) These summary statistics are two-step estimators and so their variance must reflect:

1. The first step uncertainty from estimating the counterfactual distribution.
2. The second step uncertainty from estimating the statistics using sample, rather than population, data.

The resultant analytical formulae for the standard errors are extremely complex! (Note that this is similar in spirit to the DFL/IPW estimator discussed in previous lectures, though the implementation is very different.)

In both cases, the bootstrap standard errors described in the previous section provide simple alternatives to analytical standard error estimators.

Bias correction

Many widely-used estimators are consistent but biased in finite samples. The bootstrap allows us to "bias-correct" these estimators and avoids the need for complex calculations. Consider estimating the square of the population mean, $\theta = \mu^2$. The square of the sample mean is a consistent but biased estimator, as

$$E\left[\hat{\theta} - \theta\right] = E[\bar{X}^2] - \mu^2 = \left(\mu^2 + \frac{1}{n}[E[X_i^2] - E[X_i]^2]\right) - \mu^2 = \frac{1}{n}[E[X_i^2] - E[X_i]^2]$$

(Provided the sample observations are *iid*.) The bias occurs because we cannot in general interchange the expectation operator with nonlinear functions, including polynomial transformations. In order to bias-correct this estimator, we requires unbiased estimates of each term in $\frac{1}{n}[E[X_i^2] - E[X_i]^2]$.

Alternatively, we can implement the following bootstrap algorithm:

1. Using the original data, construct an estimate of the parameter of interest, $\hat{\theta}$.
2. For b = 1 to B, where B is the total number of bootstrap samples
    a. Pick a bootstrap sample $X_1^{*b},..., X_n^{*b}$ of size n from $\hat{f}(X)$ with replacement.
    b. Calculate $\hat{\theta}_b = \bar{X}^2$ for each replication.
3. Calculate the *bootstrap estimate of the bias*:
$$Bias_{\hat{\theta}} = \left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b\right) - \hat{\theta}$$
To understand this definition, recall that the bias of an estimator is defined as the difference between its expected value and the true value of the parameter. As the bootstrap procedure treats the original sample as the population, $\hat{\theta}$ is the true value of parameter estimated by $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$. The expression above is therefore the expected value of the bootstrap estimator in repeated samples from the "population," minus the value of the parameter in the "population,"
4. Calculate the *bootstrap bias-corrected estimator*
$$\hat{\theta}_{Boot} = \hat{\theta} - Bias_{\hat{\theta}} = 2\hat{\theta} - \left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_B\right)$$

More generally, we can use Taylor series approximations to write any $\sqrt{N}$-consistent estimator (which includes almost all common parametric estimators) as
$$E\left[\hat{\theta} - \theta\right] = \frac{a_n}{n} + \frac{b_n}{n^2} + \frac{c_n}{n^3} + \cdots$$
where a_n, b_n, and c_n are bounded constants.[1] The bias corrected estimator works because it sets $a_n = 0$, ensuring that the estimator converges to its probability limit "faster" as the sample size increases.

However, this is not a panacea! The bias-corrected estimate changes the values of b_n and c_n and it may in fact have larger finite sample bias. Furthermore, the bias-corrected estimator often has a larger variance than the uncorrected estimator, so the correction may increase mean squared error if the bias is small.

In practice, bias correction is most commonly used for nonparametric estimators, which converge more slowly with the sample size. Many of these estimators are also asymptotically biased, even if they are consistent, so the bias is typically a source of greater concern than in parametric models.

[1] Note that in the example above, $a_n = E[X_i^2] - E[X_i]^2$ and $b_n = c_n = \cdots = 0$.

Inference

The simplest application of the bootstrap to statistical inference is variance estimation, described above. Here we generate a point estimate for the parameter of interest, bootstrap its standard error, construct the relevant test statistic and then compare this statistic to the critical values from the relevant reference distribution (typically the standard normal or χ^2 distributions).

Given this simplicity, it is tempting to use this as the default bootstrap choice. However, alternative bootstrap procedures may be preferable in certain situations. In particular, alternative bootstrap procedures might:

1. Permit inference procedures that do not directly rely on limiting distributions and on the critical values from these distributions.
2. Generate statistics that converge to their limiting distributions "faster" than analytical statistics. Bootstrap procedures based on such statistics are said to offer *asymptotic refinement*.

The following bootstrap procedure avoids the need to rely on critical values from a limiting distribution:

1. Using the original data, construct an estimate of the parameter of interest, $\hat{\theta}$.
2. For b = 1 to B, where B is the total number of bootstrap samples
    a. Pick a bootstrap sample $X_1^{*b},...,X_n^{*b}$ of size n from $\hat{f}(X)$ with replacement.
    b. Calculate $\hat{\theta}^{*b} = s(X^{*b})$.
3. Order the B bootstrap estimates from smallest to largest and reject the null hypothesis at a 10% significance level if θ_0 is smaller than the 5th percentile or larger than the 95th percentile of the distribution.
4. Alternatively, construct a 90% confidence interval for θ_0 as the difference between the 5th and 95th percentiles.[2]

Intuitively, this algorithm approximates the sampling distribution of the coefficient and then rejects the null hypothesis if the hypothesized value of the parameter is unlikely to have been observed under the null (i.e. if the hypothesized value is in the tail of the sampling distribution). This bootstrap algorithm does not assume that the coefficient is normally distributed, which is a necessary condition for the usual standard error-based confidence intervals to have approximately correct coverage. However, this bootstrap procedure does not offer asymptotic refinement.

To understand these points, we note that a *pivotal statistic* is a statistic whose distribution does not depend on any unknown parameters. Consider the classical normal linear regression model with a single slope parameter. The least squares estimators of the

[2] Note that in general, the distances between the critical values $\hat{\theta}_{0.05}$ and $\hat{\theta}_{0.95}$ and the original estimate will not be equal, so the confidence interval will not typically be symmetric.

parameters of the models are not pivotal, as $\hat{\beta}$ converges in probability to β_0, an unknown parameter. Similarly, the centered parameter $\hat{\beta} - \beta_0$ is not pivotal, as its asymptotic variance-covariance matrix is unknown. The studentized statistic $(\hat{\beta} - \beta_0)/\hat{\sigma}$ has a t_{n-2} distribution under the assumption of normal errors and so is pivotal.[3] Conveniently, this is the usual test statistic for the null hypothesis that $\hat{\beta} = \beta_0$, which illustrates the general point that *test statistics are often pivotal*.

What if the assumption of normal error terms is relaxed? It is still true that $(\hat{\beta} - \beta_0)/\hat{\sigma}$ has a standard normal limiting distribution and so is *asymptotically pivotal*. Under standard assumptions, the *t*-statistics used to test single hypotheses and the Wald, likelihood ratio and Lagrange multiplier statistics used to test multiple hypotheses are all asymptotically pivotal. In general, tests that apply the bootstrap to asymptotically pivotal statistics converge to the correct size at rate n, while tests relying on non-pivotal statistics converge at rate $\sqrt{n}$.[4] See Cameron & Trivedi (2005, 371-372) for a terse but clear explanation or Horowitz (2001, *Handbook of Econometrics*, section 3) for a more detailed but considerably more technical explanation.

Furthermore, *symmetric two-sided* tests that apply the bootstrap to asymptotically pivotal statistics converge even faster, at rate $\sqrt[3]{n}$. This result applies, for example, to test statistics whose asymptotic distribution is standard normal.

The following bootstrap procedure takes advantage of the faster convergence of the pivotal test statistic:

1. Using the original data, construct an estimate of the parameter of interest, $\hat{\theta}$, its standard error $\widehat{se}(\hat{\theta})$, and the test statistic $t = (\hat{\theta} - \theta_0)/se(\hat{\theta})$ The standard error estimate should be obtained from a consistent estimator that may be biased.
2. For *b = 1* to *B*, where *B* is the total number of bootstrap samples:
    a. Pick a bootstrap sample $X_1^{*b}, ..., X_n^{*b}$ of size *n* from $\hat{f}(X)$ with replacement.
    b. Calculate $\hat{\theta}_b$ and $t_b = (\hat{\theta}_b - \hat{\theta})/\widehat{se}(\hat{\theta})$.
3. Order the *B* test statistics from smallest to largest and reject the null hypothesis at a 10% significance level if *t* is outside the *bootstrap critical values* (i.e. is smaller than the 5th percentile or larger than the 95th percentile of the distribution).
4. Alternatively, construct a 90% *bootstrap confidence interval* for *t* as the difference between the 5th and 95th percentiles.

[3] Assume that the appropriate degrees of freedom correction has been incorporated into $\hat{\sigma}$.

[4] Recall that a statistic $\hat{\theta}$ "converges at rate $\sqrt{n}$" if $\hat{\theta} = \theta_0 + O(n^{-1/2})$ and the same statistic converges "converges at rate n" if $\hat{\theta} = \theta_0 + O(n)$. The latter statistic converges to its population value faster as the sample size increases, although it is not necessarily closer to this value in any given finite sample.

Implementation choices

It has already become clear that there are a range of different bootstrap procedures available. In this section we discuss several choices involved in implementing the bootstrap. You will typically face all of these choices in each application of the bootstrap: you will need to choose a test procedure, how much structure to impose, which units to resample, and the number of replications. These choices are not independent, so your choice of a test procedure, for example, may influence your choice of the number of replications to use.

Which test procedure to choose?

We have so far introduced three possible test procedures (assuming a 10% significance level for simplicity):

1. Use the bootstrap to estimate the standard error of the estimator of interest, construct a test statistic, and reject the null hypothesis if the hypothesized value is more than 1.65 standard bootstrap deviations from the estimated coefficient (page 2).
2. Use the bootstrap to estimate the sampling distribution of the estimator of interest and reject the null hypothesis if the hypothesized value is smaller than 5^{th} percentile or larger than the 95^{th} percentile of the bootstrap distribution (page 5). This is often known as the *percentile bootstrap*.
3. Use the bootstrap to estimate the sampling distribution of the test statistic and reject the null hypothesis if the hypothesized value is smaller than 5^{th} percentile or larger than the 95^{th} percentile of the bootstrap distribution (page 6). This is often known as the *percentile-t bootstrap*.

The first option has the advantage of simplicity and allows the reader to mentally construct confidence intervals for any hypothesis test in which they are interested. The second and third approaches yield results (confidence intervals, *p*-values) that are specific to a single null hypothesis and so restrict readers' ability to perform casual inference. Angrist & Pischke (2008, chapter 8.2) lean toward the first option for this reason.

However, the second and third procedures avoid the need to rely on critical values from the asymptotic distribution and the third offers asymptotic refinement. These procedures are also valid if the statistic of interest has a skewed distribution, whereas the standard error-based bootstrap procedures impose symmetric confidence intervals. Horowitz strongly argues in favor of the percentile-t for this reason (2001, *Handbook of Econometrics*, pp. 3163). Aside from theoretical attraction, asymptotic refinement may be substantively important when the sample size is small. Cameron, Gelbach & Miller (2006, *Review of Economics and Statistics*) provide a relevant example.

`[Omit the following section in the lecture.]`

It is also possible to combine bootstrap-based inference with bootstrap bias correction. The key idea is to correct the coefficient estimate for its bias and then construct a confidence interval around the bootstrap bias-corrected estimator. Note that in this

context we typically use a median-based bias correction, rather than a mean-based bias correction. Define the *median bias* as

$$z_0 = \Phi^{-1}\left(\frac{\#(\hat{\theta}^{*b} \leq \hat{\theta})}{B}\right)$$

The term inside the inverse normal CDF is the proportion of times that the bootstrap coefficient estimate is smaller than the original coefficient estimate. When the estimator is median-unbiased, this fraction equals one half and so $z_0 = 0$.

For a 90% confidence interval, define

$$p_l = \Phi(2z_0 - z_{0.95})$$
$$p_u = \Phi(2z_0 + z_{0.95})$$

Mechanically, the multiplication by 2 arises in the same way as the bias correction on pages 3-4 above. Intuitively, we can think of this as a double bias correction: one correction for the bias of the original estimator relative to the true population parameter and one correction for the bias of the bootstrap estimator relative to the original estimator.

We can now construct a bias-corrected confidence interval using the bias-corrected estimator. The lower and upper bounds of the confidence interval are given by the ${p_l}^{th}$ and ${p_u}^{th}$ percentiles of the distribution of bootstrapped parameter estimates. Note that if the estimator is median unbiased, this is the "standard" percentile bootstrap confidence interval. See Efron (1981, *Biometrika*) for the original derivation.

Efron (1987, *Journal of the American Statistical Association*) generalizes the bias-corrected bootstrap confidence intervals to *bias-corrected and accelerated bootstrap confidence intervals* (BCA). This generalizes p_l and p_u to

$$p_l = \Phi\left(z_0 + \frac{z_0 - z_{0.95}}{1 - a(z_0 - z_{0.95})}\right)$$
$$p_u = \Phi\left(z_0 + \frac{z_0 + z_{0.95}}{1 - a(z_0 + z_{0.95})}\right)$$

for a constant *a*, called the *jackknife estimate of the acceleration*. Note that this more general bootstrap estimator nests the earlier case when *a=0* and that the confidence interval collapses to zero as a goes to infinity. The value of *a* is based on the skewness of the underlying data and software packages such as Stata automatically approximate *a*. Intuitively, the BCA confidence intervals explicitly correct for potential skewness in the distribution of the parameter, whereas "standard" percentile bootstrap procedures do so more casually.

The "acceleration" Efron establishes for this estimator is equivalent to asymptotic refinement, discussed above. This is an attractive feature of BCA confidence intervals relative to "standard" percentile bootstraps. However, Efron's Monte Carlo results suggest that tests based on this bootstrap procedure require a relatively high number of replications (at least 1000 for well-behaved estimators). These confidence intervals are fairly common in applied work but receive little attention in econometric theory, which has focused more on percentile-*t* bootstraps based on pivotal statistics.

How much structure to impose?

The bootstrap algorithms described above are all *nonparametric*, in the sense that the resampling scheme does not assume that the data are drawn from any specific distribution. We can impose additional structure on the data in a variety of ways. As a rough heuristic, imposing structure will lead to narrower confidence intervals, which can be interpreted as an efficiency gain if the structure is correct and a bias otherwise.

Consider the following example adapted from Horowitz (2001, *Handbook of Econometrics*, 3185 – 3186). We wish to test the hypothesis H_0: $\beta = \beta_0$ in the model

$$Y = \alpha + \beta X + \varepsilon$$

We can implement any of the following resampling procedures:

1. Resample (Y_b, X_b) with replacement from the sample data. This is typically known as the *nonparametric bootstrap* or *pairs bootstrap*.

2. Estimate $(\hat{\alpha}, \hat{\beta})$ using the original data and construct $\hat{u} = Y - \hat{\alpha} - \hat{\beta}X$. Then in each replication, draw $(\hat{u}_b, X_b)$ with replacement from the sample data and construct $\hat{Y}_b = \hat{\alpha} + \hat{\beta}X_b + \hat{u}$. This is typically known as the *residual bootstrap*.

   A variant on this procedure keeps the original *X* matrix, resamples $\hat{u}_b$ and then constructs $\hat{Y}_b = \hat{\alpha} + \hat{\beta}X + \hat{u}_b$. Conceptually, this treats *X* as fixed, rather than a random variable, and so may be more appropriate in experiments. In practice, the two approaches seldom yield different results (Efron & Tibshirani, 1986).

3. Estimate $\hat{\alpha}$ using the original data, imposing the condition that $\beta = \beta_0$ and construct $\hat{u} = Y - \hat{\alpha} - \beta_0 X$. Then in each replication, draw $(\hat{u}_b, X_b)$ with replacement from the sample data and construct $\hat{Y}_b = \hat{\alpha} + \beta_0 X_b + \hat{u}$. This is typically known as the *residual bootstrap with the null hypothesis imposed*.

4. Construct the fitted residuals as in the second procedure and assume that they are drawn from a specific distribution, say $N(\mu, \sigma^2)$. Then use the mean and variance of the fitted residuals to estimate μ and σ^2. Finally, resample $\hat{u}$ from the $N(\hat{\mu}, \hat{\sigma}^2)$ distribution, rather than the original data and use the original *X* data in each replication. This is one example of a *parametric bootstrap*.

The "correct" procedure depends on the specific nature of the application – the bootstrap is not an excuse for avoiding thinking about the underlying model! Nonetheless, these suggestions may be useful:

1. Roughly speaking, moving down the list imposes more structure on the data, which yields efficiency gains if the structure is correct but may produce incorrect inferences. Less structure should therefore lead to more conservative inferences. If your "preferred" bootstrap imposes substantial structure on the data, it may be

useful to check whether the results are robust to using a nonparametric bootstrap instead. If not, this should be a source of concern.

2. The second version of the residual bootstrap implicitly assumes homogeneity. This is particularly problematic if there is heteroskedasticity that depends on *X*: $\sigma_i^2 = \sigma_i^2(X_i)$.

3. Monte Carlo simulations are your friend! When unsure about the correct bootstrap algorithm to apply, consider specifying the most likely data generating process, simulating some data, and applying several different bootstrap procedures to see which produces the most accurate tests.

Which units to resample?

All applications of the bootstrap discussed sampled single individuals, which is appropriate if there is no correlation between the individual error terms. If correlation is suspected (clustered sampling, experiments with treatment assigned at the cluster level, panel data, time series data, etc.), it is more appropriate to resample at the level at which errors are thought to be correlated.

Theory does not always provide clear guidance on the appropriate level of clustering. For example, when a survey samples students within classrooms with schools, is the clustering at the level of the school or classroom? The answer may be context-specific but in general, clustering at higher levels is more conservative.

In practice, the empirical literature typically uses a single-level bootstrap, which resamples clusters and takes the set of observations within a cluster as given. It may be more appropriate to use a *nested bootstrap*, which resamples clusters and then resamples individual observations within the selected clusters.

See Cameron, Gelbach & Miller (2008, *Review of Economics and Statistics*) for a detailed discussion of bootstrap procedures for clustered data. Angrist & Pischke (2008, chapter 8.2) and Bertrand, Duflo & Mullainathan (2004, *Quarterly Journal of Economics*) are also useful references. There may be some additional complications when resampling clusters and using cluster fixed effects – see Kezdi (2004, *Hungarian Statistical Review*).

How many replications is "enough?"

There is unfortunately no rote answer to this question. Speaking very generally, tests based on the percentile bootstraps require getting the tail probabilities "right," and so typically require more replications than tests based on standard error bootstraps. However, with sufficiently many replications, percentile-based bootstraps are more robust to a single extreme outlier than standard error-based bootstraps.

Most referees are satisfied by 500 replications for standard error-based bootstraps and 1000 for percentile-based bootstraps, though more may be better in the latter case. You should increase the number of replications when estimating parameters that are in some sense "unusual," such as tail probabilities. Remember that you bear the fixed cost of writing code for a bootstrap but that the computer bears the entire variable cost…

There are a handful of more formal treatments in the literature. Andrews & Buchinsky (2000, *Econometrica*) derive the level of accuracy as a function of the number of replications in several specific settings. Davidson & McKinnon (2000, *Econometric Reviews*) derive the level of power as a function of the number of replications. Johnston & DiNardo (available online shortly, section 12.3) provide a very clear illustrative example.

How to bootstrap a two-step estimator?

The bootstrap is frequently used to estimate the distribution of two-step estimators, including those used in propensity score reweighting methods (DiNardo, Fortin & Lemiuex, 1996, *Econometrica*; Hirano, Imbens & Ridder, 2003, *Econometrica*), the classical normal selection correction model (Heckman, 1979, *Econometrica*), and two-stage least squares (Angrist, Imbens & Rubin, 1996, *Journal of the American Statistical Association*). Many papers also use simple generated regressors such as sample or cluster means.

In all cases, **both steps** should be implemented in each iteration of the bootstrap. Taking the first stage estimates as given and implementing only the second step can substantially understate the variance of the estimator.

Failures

Cases in which the bootstrap "fails" can be placed into three broad categories. In the first category, the researcher misspecifies the data generating process and so makes inappropriate implementation choices. This problem can often be avoided by application of more suitable bootstrap procedures. In the second category, the estimator of interest is an insufficiently "smooth" function of the data (defined below) and so bootstrap procedures do not yield consistent estimates of its sampling distribution. Using a different implementation of the bootstrap will not avoid this problem. In the third category, the hypothesized value of a parameter is at the boundary of the set of possible parameter values. It is sometimes possible to adapt the implementation of the bootstrap estimator to avoid this problem.

Inappropriate implementation choices

Most poor implementation choices arise because researchers fail to take into account heteroscedasticity or serial correlation in their data. We noted on page 13 that the residual bootstrap is inappropriate when the residuals from a regression model are heteroscedastic. Similarly, we noted on page 14 that a bootstrap algorithm that resamples individual units

is inappropriate if data has a clustered structure and there is serial correlation within clusters.

In both cases, bootstrap procedures that fail to take these features of the data into account tend to be "liberal," in the sense that tests reject null hypotheses are less often than their nominal level. The serial correlation problem may be addressed by resampling clusters, rather than individual units. The heteroscedasticity problem may be addressed by using nonparametric bootstrap procedures that resample the original data, rather than fitted residuals. Alternatively, we might interpret these features of the data as indications of model misspecification and try to improve the model.

Wu (1986, *Annals of Statistics*) and Liu (1988, *Annals of Statistics*) propose using a *wild residual bootstrap* for heteroscedastic data. This procedure replaces the fitted residual $\hat{u}_i$ for each observation *i* with

$$\begin{aligned} -0.618\hat{u}_i \quad & with\ probability\ 0.7236 \\ 1.618\hat{u}_i \quad & with\ probability\ 0.2764 \end{aligned}$$

Horowitz (1997, *Advances in Economics and Econometrics*) shows that this estimator performs very well in Monte Carlo simulations and Cameron, Gelbach & Miller (2006, *Review of Economics and Statistics*) recommend its use for clustered data when the number of clusters is small. The latter paper's Monte Carlo results suggest that the scaling factors can be replaced by (-1,1) and the probabilities by (0.5,0.5) although the original scaling factors and probability weights have some desirable theoretical properties.

Bertrand, Duflo & Mullainathan (2004, *Quarterly Journal of Economics*) also note that serial correlation through time in panel data may be a problem. They emphasize that many microeconometric applications use panel data from short time frames over which data is highly correlated (for example, annual data on US states). Parametric corrections, such as modeling autoregressive processes in the data, perform poorly in their Monte Carlo simulations, relative to resampling state clusters, rather than state-year clusters.

The same problem applies in time series data, where we typically worry about serial correlation in the residuals. We can try to address this by modeling the temporal dependence in the errors or by using a *moving block bootstrap* that resamples non-overlapping blocks of *t* years from a sample of *T*<*t* data points. See Götze and Künsch (1996, *Annals of Statistics*) for a very technical explanation of why and under what conditions this bootstrap "works." In particular, note that the size of the blocks should increase with the sample size.

Nonsmooth estimators

To understand the problem with bootstrapping the sampling distribution of non-smooth statistics, we delve briefly into some formal theory. Define *t* as the statistic of interest, G_0 as the true distribution (CDF) of the statistic of interest, G_n as the limit of its sampling distribution as the sample size *n* goes to infinity, F_0 as the true distribution of the data,

and F_n as the sample distribution of the data. For most parametric statistics, the true distribution equals the limit distribution plus some error that converges to zero at rate $\sqrt{n}$:

$$G_0(t, F_0) = G_n(t, F_0) + O(n^{-1/2})$$

or, more precisely:[5]

$$G_0(t, F_0) = G_n(f, F_0) + \frac{g_1(t, F_0)}{n^{1/2}} + O(n^{-1})$$

We can write a similar expansion based on the bootstrap estimator of the data distribution:

$$G_0(t, F_n) = G_n(f, F_n) + \frac{g_1(t, F_n)}{n^{1/2}} + O(n^{-1})$$

Using these expansions, we can write the deviation of the bootstrap-based estimator of the statistic's sampling distribution from the statistic's true sampling distribution as

$$\begin{aligned} G_n(t, F_n) - G_n(t, F_0) &= [G_0(f, F_n) - G_0(f, F_0)] + \left[\frac{g_1(t, F_n) - g_1(t, F_0)}{n^{1/2}}\right] + O(n^{-1}) \\ &= [G_0(f, F_n) - G_0(f, F_0)] + O(n^{-1}) \end{aligned}$$

where the second equality follows because $g_1(t, F_n) - g_1(t, F_0) = O\left(n^{-1/2}\right)$ under general conditions. The expression on the right-hand side, although complex, yields several important insights into the operation of the bootstrap:

1. For asymptotically pivotal statistics, the distribution of the statistic G_0 does not depend on unknown parameters of the distribution of the data, F_0, so $G_0(f, F_n) = G_0(f, F_0)$ and the leading term in the expression is zero. This illustrates why asymptotically pivotal statistics offer asymptotic refinement.
2. The sample distribution of the data, F_n, converges to the population distribution, F_0, as the size of the original sample, *n*, approaches infinity. This illustrates why bootstrap approximations to sampling distributions improve with the size of the original sample.
3. The entire expression converges to zero as the sample distribution F_n converges to the population distribution F_0 only if G_n is a "smooth" functional. If this functional has discontinuous jumps, then it is possible that as the sample and population distributions become close, there are sharp jumps in the sampling distribution of the statistic of interest.[6]

Estimators whose non-smoothness causes the bootstrap to fail include nearest-neighbor matching (Abadie & Imbens, 2006, *Econometrica*) and perhaps the maximum score

[5] This expression uses an *Edgeworth expansion*, so the functions g_1 and g_2 are based on the cumulants of G_0.

[6] Most theoretical papers that wish to establish the validity of the bootstrap do so by showing that the estimator being bootstrapped is a Hadamard differentiable functional of the data distribution. See Chernozhukov, Fernandez-Val and Melly (2009, mimeo) and Firpo (2010, mimeo) for examples. As a practitioner, it is typically sufficient to note that most estimators whose descriptions "sound smooth" satisfy this condition; trying to develop a more rigorous understanding may entail a large fixed cost.

estimator (Manski, 1975, *Journal of Econometrics*).[7] In the former case, the smoothness condition fails because there are discontinuous switches from one nearest neighbor to another with small changes in the data. In the latter case, Manski proposes a semiparametric estimator for the binary choice model based on the empirical distribution function, so small changes in the data can discontinuously switch an observation from a probability of zero to a probability of one.

In both cases, there exist "smoother" alternative estimators for which the bootstrap is unlikely to fail. Nearest neighbor matching can be replaced by kernel matching or propensity score reweighting and the empirical distribution function in the maximum score estimator can be smoothed by density-weighted estimators (Horowitz, 1992, *Econometrica*).

Boundary problems

The bootstrap will have troubles in cases where the population value of a parameter is on the boundary of the set of possible parameter values. For example, suppose that you want to test the null hypothesis that the impact variance equals zero. This example is considered in Appendix E to Heckman, Smith and Clements (1997, *Review of Economic Studies*).

To do this, you first note the result that the lower bound on the impact variance is given by distribution of impacts implied by the quantile treatment effects estimator. Then, you might naively proceed to estimate a bootstrap confidence interval on the variance of this distribution and check to see if it contains zero.

The trouble here is that it will never contain zero unless both the null holds and there is no error term. The reason is that the variance in every bootstrap sample must be non-zero due to random variation combined with the fact that the variance squares all of the estimated quantile treatment effects. The solution is to bootstrap under the null hypothesis using only the control group observations.

Note that this problem also arises when using conventional analytical estimators. The most common sufficient conditions for consistency of M-estimators (which include least squares, maximum likelihood, quantile regression, and generalized method of moments) assume that the true value of the parameter is not on the boundary of the parameter space (Newey & McFadden, 1994, *Handbook of Econometrics*).

Horowitz (2001, *Handbook of Econometrics*, sections 2.1 and 4.5) discusses several other cases in which the bootstrap may be valid but still performs very poorly, such as (1) weak instruments, (2) variance estimators that themselves have high variances, (3) estimators of sample maxima and minima, and (4) estimators of the mean for a Cauchy population. As a rough heuristic, the bootstrap should be used with caution whenever its analytical

[7] Horowitz (2001, section 4.3.2) notes that there does not appear to be a formal proof of the validity or failure of the bootstrap for the maximum score estimator. This may in part reflect its very complicated, non-normal asymptotic distribution, which makes analysis (or use!) of this estimator challenging.

equivalent would be used with caution. The bootstrap is neither a magic bullet nor a substitute for careful thought!

Alternatives

We briefly discuss two procedures that are in some instances appropriate alternatives to the bootstrap. Before proceeding, it is worth recalling that one original motivation for the bootstrap was to avoid the need for complicated analytical expressions. In cases where the bootstrap fails, analytical estimators may sometimes be viable alternatives.

The jackknife

The jackknife is a deterministic resampling procedure that takes *n* samples of size *n-1*, obtained by systematically dropping one observation from the original data. Define $\hat{\theta}_{-i}$ as the estimate of θ omitting observation *i*. Then the *jackknife estimate of the standard error* of θ is defined as

$$\widehat{se}_J(\hat{\theta}) = \left[\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{-i} - \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}\right)^2\right]^{1/2}$$

and a *bias-corrected jackknife estimate* of θ can be constructed by subtracting the *jackknife estimate of the bias*, $(n-1)\left(n^{-1}\sum_{i=1}^{n}\hat{\theta}_{-i} - \hat{\theta}\right)$, from $\hat{\theta}$. These estimators are very similar to their bootstrap equivalents, with the rescaling factor $n-1$ reflecting the difference between the size of the jackknife samples and the original data.

The jackknife predates the bootstrap by several decades and is still in widespread use in the survey sampling literature. Efron & Tibshirani (1993) derive the relationship between the jackknife and bootstrap under certain conditions. The jackknife is less widely used in the econometrics literature, though it has appeared in work on topics such as weak instruments (Phillips & Haile, 1997, *International Economic Review*), nonlinear panel models (Hahn & Newey, 2004, *Econometrica*), model averaging (Hansen & Racine, forthcoming, *Journal of Econometrics*), and semiparametric two-step estimation (Cattaneo *et al*., 2011, mimeo).

Subsampling

Subsampling is a variant on the bootstrap in which samples of size *m* are drawn from the original data, for *m* considerably smaller than *n*. The samples may be drawn with or without replacement. Horowitz (2001, *Handbook of Econometrics*) refers to these as *replacement subsampling* and *non-replacement subsampling* respectively, though there does not appear to be a uniform terminology in the literature. Replacement subsampling is mechanically similar to bootstrapping but is interpreted as drawing random samples from the population, rather than from the original data, and so may be consistent when the bootstrap is not. It yields consistent estimators of the sampling distribution of some statistics for which the bootstrap does not, including estimators of boundary parameters.

Non-replacement subsampling is mechanically more complex, as takes all possible non-replacement subsets of size *m* from the sample. This is computationally cumbersome as the number of subsets, $\binom{n}{m}$, is very large for even moderately sized samples. However, non-replacement subsampling can consistently estimate the sampling distribution of statistics under *very* weak conditions. Intuitively, this works because for $m \ll n$, the subsampling estimator, $\hat{\theta}_s$, has a far larger sampling error than $\hat{\theta}$. Hence, the sampling distribution of $\hat{\theta}_s - \hat{\theta}$ provides a good approximation to the sampling distribution of $\hat{\theta}_s - \theta$ (after both differences have been appropriately rescaled).

Subsampling techniques have an important limitation. Statistics computed using subsampling are based on far smaller samples than those computed using bootstrapping. They thus tend to be less "accurate," in the sense that $G_n(t, F_n) - G_n(t, F_0)$ converges more slowly for subsampling estimators. Perhaps for this reason, and because the bootstrap is consistent for most widely used estimators in economics, subsampling is very uncommon in applied work. Horowitz (2001, *Handbook of Econometrics*, section 2.2) provides a brief overview. Politis, Romano & Wolf (1999) provide a considerably more detailed discussion of subsampling methods.

Bootstrapping in Stata

The bootstrap command

Stata's `bootstrap` (or `bs`) command does bootstrapping. The basic syntax is:

```
bootstrap "command" exp_list, reps(#) dots
```

where "`exp_list`" is the list of statistics saved by the estimation command that the bootstrap command should save for each bootstrap sample, and then display bootstrap standard errors for.

The `reps(#)` option indicates the number of bootstrap replications; the default here is 50, which will be too low in many circumstances, as discussed earlier.

The `dots` option causes Stata to display a period (dot) every time it finishes with another bootstrap sample. As noted in the manual, this provides entertainment as you await the completion of your bootstrap replications.

There are many other options as well. You can save the bootstrap estimates to a file for later reuse and change many other aspects of what is done and how the results are displayed. This command can also accommodate stratified or clustered resampling.

Note that this command can only be used to bootstrap Stata commands that are implemented with a single line of code. This works for "canned" two-step estimators such as `heckman` (for the bivariate normal selection model) and `ivregress` for two-

stage least squares). However, it will not work for two-step estimators that need to be manually implemented.

Sample command

An example of executing the bootstrap in Stata is given by:

```
bs "psmatch2 nsw, pscore(pcpsdwc) outcome(re78)" "r(att)",
nowarn reps(250);
```

The "bca" option must be explicitly specified in order to obtain the bias corrected and accelerated bootstrap confidence interval; this method is optional presumably because it takes a while to calculate.

Practical concerns

First, you can speed up Stata's bootstrap command somewhat by dropping variables from the data set that are not used in whatever command is being bootstrapped.

Second, you should remove all the observations in the data set that will not be used in the estimation. This includes, for example, observations with values missing due to item non-response. Stata will sometimes do this automatically but is unlikely to do so for an unofficial Stata command (e.g., psmatch2).

Stata counts the number of observations in the data at the start of the bootstrap procedure and uses this number to determine how many observations to sample. If the data contain a lot of observations that cannot be used, then the bootstrap sample sizes will bounce around and will in general not match the sample size actually used in the analysis, both of which are bad things.