ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

STATISTICAL INFERENCE:

PART 3, THE BOOTSTRAP

Randomization tests are useful for testing a null hypothesis without making parametric assumptions. We may take the nonparametric route because we want to be robust to distributional assumptions, or we may take the nonparametric route because we cannot (or do not want to) calculate the finite sample (or even asymptotic) properties of our estimator.[1] But what if we want a resampling based procedure that can produce confidence intervals rather than just testing the null hypothesis? Then we turn to bootstrapping.

1 Bootstrapping the Mean

We begin by considering a simple example in which we want to estimate the variance of the sample mean, $\hat{\theta} = \bar{y}$. We view each observation, y_i, as a random draw from a larger population. One way to estimate the variance of $\bar{y}$ is to assume that $y_i \sim N(\theta, \sigma^2)$, in which case we can show that $\bar{y} \sim N(\theta, \sigma^2/N)$. Alternatively, we can relax the normality assumption and apply the Central Limit Theorem and the Law of Large Numbers to show that $\sqrt{N}(\bar{y} - \theta) \sim N(0, \sigma^2)$ asymptotically. Suppose, however, that we did not know these things (which can be the case with more exotic estimators). How might we estimate the variance of $\hat{\theta} = \bar{y}$?

One way would be to use the sample analog of the variance. No, not the sample analog $\widehat{\text{Var}}(\hat{\theta}) = \widehat{\text{Var}}(\bar{y}) = (1/N)\sum(y_i - \bar{y})^2/N$, which still relies on the formula $\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B) + 2 \cdot \text{Cov}(A, B)$. Rather, we could randomly draw S samples of size N from

[1] Computing time is cheap and getting cheaper; human time is expensive and not getting cheaper.

the population and compute S estimates of $\hat{\theta}_s = \bar{y}_s$ for $s = 1, ..., S$. Then estimate

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{S-1}\sum_{s=1}^{S}(\hat{\theta}_s - \bar{\hat{\theta}})^2$$

where $\bar{\hat{\theta}} = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}_s$. Note that this is the conventional estimator we use for the standard deviation of a random variable; the difference is that the unit of observation is now the sample rather than the individual observation.

In practice, of course, we only have one sample at our disposal, not S samples. Nevertheless, we can estimate the population distribution of y_i using the empirical distribution of y_i in our sample. We do this by randomly drawing N observations from our data set with replacement.[2] Each of these draws of N observations constitutes a single bootstrap sample, b. For a given bootstrap sample, we calculate the mean of the bootstrapped observations, $\bar{y}_b$. Repeating this procedure B times produces B estimates of the statistic $\hat{\theta}_b = \bar{y}_b$, $b = 1, ..., B$. We then estimate the variance of $\hat{\theta} = \bar{y}$ as

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\hat{\theta}})^2$$

where $\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$. This is an example of a nonparametric bootstrap; it is nonparametric in that it makes no distributional assumptions, though it still relies on independence between observations.

We have motivated the bootstrap as a method of estimating the variance of an estimator when no analytic estimate is available. In practice, that is what it is generally used for. Nevertheless, for asymptotically pivotal test statistics, the bootstrap can also be used to improve upon the first-order asymptotic approximation of $\text{Var}(\hat{\theta})$.[3] Applied researchers rarely

[2] After reading through the procedure, it should be obvious why we sample with replacement. If we sampled without replacement, $\hat{\theta}_b$ would be identical for every bootstrap sample.

[3] An asymptotically pivotal test statistic is one whose asymptotic distribution does not depend on any unknown parameters. For example, $\hat{\theta} = \bar{y}$ is not asymptotically pivotal because, even assuming the null hypothesis $E[y_i] = \theta_0$, it still depends on the unknown parameter σ. The t-statistic, $t = (\hat{\theta} - \theta_0)/s_{\hat{\theta}}$, however, is asymptotically pivotal.

use it for that purpose, however, and, with the exception of the clustered case, we will not devote much discussion to it here.

2 General Bootstrapping Procedure

For a general sample containing observations $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_N}$, suppose that we are interested in estimating the distribution of some statistic $\hat{\theta}(\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_N})$. A general bootstrapping procedure is implemented as follows:

1. Using the original sample $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_N}$, draw a bootstrap sample with replacement using one of the methods discussed in Sections 2.1, 2.2, or 3. Call this sample $\mathbf{w_1^*}, \mathbf{w_2^*}, ..., \mathbf{w_N^*}$.
2. Compute the statistic of interest, $\hat{\theta}$, using the bootstrap sample – call the resulting estimate $\hat{\theta}^*$. Note that $\hat{\theta}$ could be a coefficient, a standard error, or a test statistic.
3. Repeat the first two steps B times, collecting B iterations of the statistic, $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$.

Once you have generated the B bootstrapped values of the statistic, there are a variety of things you can do with them. The most likely candidates are to compute the sample variance of the statistic and/or to construct confidence intervals for the statistic's estimand. To compute the bootstrapped sample variance of the statistic, use the formula that we saw in Section 1:

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2$$

To compute a bootstrapped 95% confidence interval for θ, find the 2.5 and 97.5 percentiles of $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$ and use those two values as the lower and upper bounds of the confidence intervals. To estimate the distribution of the t-statistic for some estimator $\hat{\beta}$, you could let $\hat{\theta}^* = (\hat{\beta}^* - \hat{\beta})/s_{\hat{\beta}}^*$.[4] You could then define a rejection region for $\hat{\theta} = t_{\hat{\beta}}$ from the original data

[4] $s_{\hat{\beta}}^*$ is calculated the standard way, but using the bootstrap sample data instead of the original data.

using the 95th percentile of $|\hat{\theta}_1^*|, ..., |\hat{\theta}_B^*|$ (i.e., reject if $|\hat{\theta}|$ is greater than the 95th percentile of $|\hat{\theta}_1^*|, ..., |\hat{\theta}_B^*|$). Note that $\hat{\theta}^*$ is centered at its expectation in the empirical distribution so we are effectively bootstrapping the t-statistic's distribution under the null hypothesis. The advantage of bootstrapping the coefficient's t-statistic rather than the coefficient itself ($\hat{\beta}$) is that we may gain the asymptotic refinements mentioned in Section 1.

One issue is how large to set B. In most cases, several hundred iterations is sufficient, but with cheap computing time you might as well do several thousand unless your estimator is very computationally intensive.

2.1 The Paired Bootstrap

We defined a general bootstrapping procedure in Section 2, but we did not specify how to actually generate the bootstrap samples. The first, and most common, method that we consider is the nonparametric bootstrap, also known as the paired bootstrap.

Suppose that the data $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_N}$ consist of N pairs of observations, $\mathbf{w_i} = (y_i, \mathbf{x_i})$, where y_i is the dependent variable and $\mathbf{x_i}$ contains the explanatory variables. The paired bootstrap draws pairs, $(y_i^*, \mathbf{x_i^*})$ from the empirical distribution of $\mathbf{w_i}$ with replacement. In other words, for any given draw, both y_i^* and $\mathbf{x_i^*}$ come from the same observation. Thus the relationship between $\mathbf{x}^*$ and y^* is determined by the data rather than by any parametric assumptions (hence the name nonparametric bootstrap). Note that this method can easily be applied to nonlinear estimators as well as linear estimators.[5]

Although the paired bootstrap is nonparametric, it still requires the assumption of independence between observations. It randomly samples from the empirical distribution – if this random sampling assumption is unjustified, then the bootstrap confidence intervals may be too narrow.

[5]Bootstrapping wouldn't be that useful if we couldn't apply it to nonlinear estimators – we already know how to calculate the standard errors for linear estimators.

2.2 The Residual and Parametric Bootstraps

If we are willing to make additional assumptions about the data generating process, we can improve the approximation that the bootstrap provides. For example, consider a regression model of the form $y_i = g(\mathbf{x_i}, \beta) + \varepsilon_i$. After estimating $\hat{\beta}$, we can use this estimate to form the residuals, $\hat{\varepsilon}_i = y_i - g(\mathbf{x_i}, \hat{\beta})$. Note that $g(.)$ could be linear (e.g., ordinary least squares) or nonlinear (e.g., nonlinear least squares). We can then perform a residual bootstrap by constructing a sample consisting of $(y_1^*, \mathbf{x_1}), ..., (y_N^*, \mathbf{x_N})$, where $y_i^* = g(\mathbf{x_i}, \hat{\beta}) + \hat{\varepsilon}_i^*$, and $\hat{\varepsilon}_i^*$ is resampled with replacement from $\hat{\varepsilon}_1, .., \hat{\varepsilon}_N$. We call this the residual bootstrap because it resamples residuals, which are randomly assigned to some $\mathbf{x_i}$ and used to construct y_1^*.[6]

The benefit of residual bootstrapping is a potentially improved approximation. The drawback, however, is a loss of robustness. For example, suppose that there is heteroskedasticity. Because the residual bootstrap randomly assigns $\hat{\varepsilon}_i^*$ to $\mathbf{x_i}$, there is no heteroskedasticity in the bootstrap samples ($\mathbf{x_i}$ cannot predict the variance of $\hat{\varepsilon}_i^*$). Hence residual bootstrapping is not robust to heteroskedasticity.

The parametric bootstrap goes even further than the residual bootstrap in incorporating a priori information. Suppose we assume that the conditional distribution of y, $y_i \sim f(\mathbf{x_i}, \theta)$, is known up to the parameter θ. We estimate $\hat{\theta}$ via maximum likelihood estimation (or some other method). We perform a parametric bootstrap by constructing a sample consisting of $(y_1^*, \mathbf{x_1}), ..., (y_N^*, \mathbf{x_N})$, where y_i^* is randomly generated by the distribution $f(\mathbf{x_i}, \hat{\theta})$.[7]

A simple example of the parametric bootstrap may make thing clearer. Consider the ordinary linear regression model combined with the assumption that $\varepsilon_i \sim N(0, \sigma^2)$. In this case, $y_i \sim N(\mathbf{x_i}\beta, \sigma^2)$. Using the OLS estimates of β and σ^2, we randomly draw a y_i^* from the $N(\mathbf{x}\hat{\beta}, \hat{\sigma}^2)$ distribution for each $\mathbf{x_i}$. Our bootstrap sample is then $(y_1^*, \mathbf{x_1}), ..., (y_N^*, \mathbf{x_N})$. We repeat this B times to construct B bootstrap samples.

The advantages and disadvantages of the parametric bootstrap are similar to those of

[6] Sometimes $\mathbf{x_i}$ is also resampled before $\hat{\varepsilon}_i$ is resampled – it shouldn't make a big difference either way.
[7] Alternatively, we could also resample $\mathbf{x_i}$ before generating the y_i^*.

the residual bootstrap (better approximation vs. less robustness), only amplified. Since we are generally more concerned about robustness than about (generally marginal) efficiency improvements, the residual and parametric bootstraps are not often used by applied researchers.

3 The Bootstrap and Clustering

All of the bootstrap procedures discussed above assume independence between observations. Often we would like to apply the bootstrap in the context of clustered data, however, because for non-standard estimators it may be inconvenient or infeasible to compute asymptotic standard errors that are cluster robust. One possibility is to use the cluster bootstrap, also known as the block bootstrap. The essential idea here is that we resample at the cluster level rather than the observation level.

Suppose that we have G clusters, each containing T observations, and that G is relatively large. We represent the data as $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_G}$, where $\mathbf{w_g} = [(y_{g1}, \mathbf{x_{g1}}), .., (y_{gT}, \mathbf{x_{gT}})]$. The cluster bootstrap draws G clusters from the empirical distribution of $\mathbf{w_g}$ with replacement, where each cluster consists of the elements $\mathbf{w_g^*} = [(y_{g1}, \mathbf{x_{g1}}), .., (y_{gT}, \mathbf{x_{gT}})]$. Thus the procedure is similar to the paired bootstrap except that we are resampling clusters instead of resampling individual observations. We repeat this procedure B times to obtain B bootstrap samples.

We saw in previous lectures that cluster robust standard errors can be inaccurate when G is small. Can the cluster bootstrap provide a better approximation of the variance than cluster robust standard errors when G is small? Cameron, Gelbach, and Miller (2008) argue yes. They suggest a cluster bootstrap using the t-statistic, $\hat{\theta}^* = (\hat{\beta}^* - \hat{\beta})/s^*_{\hat{\beta}}$ rather than the coefficient itself – the t-statistic is attractive because it is asymptotically pivotal, enabling the possibility of asymptotic refinements. Note that $s^*_{\hat{\beta}}$ now corresponds to the cluster robust standard error for β rather than the conventional OLS standard error for β. The basic procedure is to construct B bootstrap samples using the cluster bootstrap, collect $\hat{\theta}^*_1, ..., \hat{\theta}^*_B$,

and reject H_0 if the absolute value of the t-statistic from the original data is greater than the 95th percentile of $|\hat{\theta}_1^*|, ..., |\hat{\theta}_B^*|$.

CGM (2008) suggest a further refinement to the procedure above, however, based on the wild bootstrap. The wild bootstrap is similar to the residual bootstrap in that it begins with the calculation of $\hat{\varepsilon}_i = y_i - g(\mathbf{x_i}, \hat{\beta})$. We then construct a sample consisting of $(y_1^*, \mathbf{x_1}), ..., (y_N^*, \mathbf{x_N})$. Instead of resampling from $\hat{\varepsilon}_1, .., \hat{\varepsilon}_N$ to get $\hat{\varepsilon}_i^*$, however, we set $\hat{\varepsilon}_i^*$ equal to $\hat{\varepsilon}_i$ with 50% probability or $-\hat{\varepsilon}_i$ with 50% probability.[8] That is to say, each $\mathbf{x_i}$ is assigned the residual from observation i with 50% probability or the negative of the residual from observation i with 50% probability. We then construct y_i^* as $y_i^* = g(\mathbf{x_i}, \hat{\beta}) + \hat{\varepsilon}_i^*$. You might think that the wild bootstrap would perform poorly because each bootstrapped observation is drawn from a distribution with only two points of support, but in fact it performs quite well. It is also robust to heteroskedasticity (unlike the residual bootstrap) because the original relationship between $\mathbf{x_i}$ and the variance of the residual is maintained.

In the context of clustering, we modify the wild bootstrap to apply to clusters rather than individual observations. The wild cluster bootstrap for the t-statistic, $\hat{\theta}^* = (\hat{\beta}^* - \hat{\beta})/s_{\hat{\beta}}^*$ is implemented as follows.

1. Estimate the OLS estimator $\hat{\beta}$ and use it to construct the residuals $\hat{\varepsilon}_g$ $(g = 1, ..., G)$. Note that $\hat{\varepsilon}_g$ is a $T \times 1$ vector containing all the residuals for cluster g.

2. For each cluster g, generate $\hat{\varepsilon}_g^* = \hat{\varepsilon}_g$ with 50% probability or $\hat{\varepsilon}_g^* = -\hat{\varepsilon}_g$ with 50% probability. Set $y_g^* = \mathbf{x_g}\hat{\beta} + \hat{\varepsilon}_g^*$. Doing this for each cluster produces a bootstrap sample consisting of $(y_1^*, \mathbf{x_1}), ..., (y_G^*, \mathbf{x_G})$. Using this sample, compute $\hat{\theta}^* = (\hat{\beta}^* - \hat{\beta})/s_{\hat{\beta}}^*$. Note again that $s_{\hat{\beta}}^*$ now corresponds to the cluster robust standard error for β rather than the conventional OLS standard error for β.

3. Repeat the second step B times to construct B bootstrap samples. Reject H_0 if the absolute value of the t-statistic from the original data is greater than the 95th percentile

[8] In Cameron and Trivedi they suggest assigning $1.618 \cdot \hat{\varepsilon}_i$ with 27.64% probability or $-0.618 \cdot \hat{\varepsilon}_i$ with 72.36% probability, but CGM (2008) suggest the simpler 50/50 probabilities.

of $|\hat{\theta}_1^*|, ..., |\hat{\theta}_B^*|$.

Using both simulated and real data, CGM find that the wild cluster bootstrap outperforms other cluster bootstraps, particularly when G is 10 or less.

4 Bootstrapping Vs. Randomization Tests

Since both methods are based on resampling, it is tempting to conclude that bootstrapping and randomization tests are the same thing. There are, however, fundamental differences between the two procedures. Philosophically, randomization tests are based off of the insight that the distribution of the test statistic arises from the random assignment procedure rather than from resampling within a set population. As a result, randomization tests can only be used to test the null hypothesis, while bootstrapping can be used to construct confidence intervals.

Randomization tests are also, in general, less parametric than bootstraps. The parametric bootstrap depends upon distributional assumptions, while the residual bootstrap depends upon homoskedasticity. Even the paired bootstrap depends on a zero serial correlation assumption.[9] Randomization tests depend upon modeling the distribution (i.e., assignment) of the treatment correctly (e.g., is it or is it not serially correlated), but this is often much clearer to the researcher than assumptions about unobserved residuals. And while the bootstrap can be modified to accommodate clustering, as in the wild cluster bootstrap, it cannot easily be modified to accommodate a challenging scenario such as the one discussed in Aker (2010) (see lecture notes on randomization tests).

[9]The bootstrap also generally depends on the assumption that the estimator is smooth.