Lecture 10
Bootstrap I

1 Introduction

Bootstrap is an alternative to asymptotic approximation for carrying out inference. The idea is to mimic the variation from drawing different samples from a population by the variation from redrawing samples from a sample. The former variation is the object of interest but is impossible to observe because one only has one sample. Bootstrap uses the latter variation to approximate the former. The name comes from the common English phrase "bootstrap" which alludes to "pulling oneself over the fence by pulling on ones own bootstrap", and means solving a problem without external help. Here it means sampling from an existing sample itself without using external samples.

To fix ideas, consider the location model: $X = \mu + \varepsilon$, where $E(\varepsilon) = 0$. The parameter of interest is μ. With an i.i.d. sample of $X : \{X_1, ..., X_n\}$, consider the standard estimator for μ: $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. To assess the accuracy of the estimator, we would like to know the distributiono $\bar{X}_n$. To know the exact distribution, we would need to obtain many more samples of size n, say $\{X_{1,k}, ..., X_{n,k}\}$ for $k = 1, 2, 3, ...$, calculate the sample average for every sample and plot the density. This (a) is not possible and (b) would be pointless if were possible because in that case μ would be known for certain.

The idea of bootstrap (Efron, 1979) is to treat the sample $\{X_1, ..., X_n\}$ as the population, and draw (with replacement) samples of size n from this "population": $\{X_1^*, ..., X_n^*\}$, calculate the bootstrap sample mean: $\bar{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$, and use the conditional distribution of $\bar{X}_n^*$ to approximate the distribution of $\bar{X}_n$. This approximation should be good if the sample $\{X_1^*, ..., X_n^*\}$ is close to the population.

The procedure of drawing a sample of size n treating the original sample as the population is called the nonparametric i.i.d. bootstrap and was proposed first in Efron (1979). Nowadays, the term "bootstrap" actually contains a large collection of self-sampling (or re-sampling) methods including the nonparametric i.i.d. bootstrap, the parametric bootstrap (for maximum likelihood models), the residual bootstrap and the wild bootstrap (both for regression models). The nonparametric i.i.d. bootstrap itself is generalized to the weighted bootstrap. An approximated bootstrap method similar to the weighted bootstrap is the score bootstrap and was proposed very recently. For this lecture, we focus on the nonparametric i.i.d. bootstrap, which is the default bootstrap method used in applications. Brief introduction of the other bootstrap methods are given along the way.

Bootstrap methods are used in place of asymptotic approximation for a number of reasons, including:

1. to gain better approximation accuracy than the asymptotic distribution, or in other words, to gain "high-order" refinement. More accurate approximation means more accurate size for a test or more accurate coverage probability for a confidence interval.

2. to bias correct an estimator. Most nonlinear estimators, though maybe consistent, are biased. Bootstrap method can sometimes estimate the bias up to some asymptotic order. The bootstrap bias estimate can then be sutracted from the original parameter estimator to give rise to a less biased estimator.

3. to avoid the calculation of the estimated asymptotic variance, or in other words, to obtain standard error without programming out the asymptotic variance formula. As it turns out, this is perhaps the most common reason to use bootstrap in applied work in economics because many economic models are so complicated that computing the standard error analytically using asymptotic distribution is very difficult.

4. sometimes, bootstrap is proposed as a method for inference for a new estimator when asymptotic distribution of the estimator has not been developed. This however is dangerous practice because bootstrap is not a panacea and fail in most cases where asymptotic approximation fails. It may also fail when asymptotic approximation succeeds. For the latter case, a notable example is the nearest neighbor matching estimator used in treatment effect. The estimator is shown to be $\sqrt{n}$-asymptotically normal, with consistently estimable asymptotic variance. However, Abadie and Imbens (2008) show that bootstrap standard errors is not correct – it some times is too large and sometimes too small.

Though bootstrap is an alternative to asymptotic approximation, it is an not an alternative to asymptotic theory. Its own justification relies on asymptotic theory. The justfication for the purposes 1-3 above has been developed. We discuss them one by one. But before doing so, we discuss bootstrap simulation first, which is the way to plot the conditional distribution of the bootstrap statistic in practice.

2 Bootstrap Simulation

Because the bootstrap samples are drawn from an artificial population (i.e. the original sample), conditional on the original sample, in principle, any statistic computed using the bootstrap sample is known. Thus, the conditional mean, variance, etc can in principle be calculated as we did for the location model above. However, in most realistic cases, the statistics has nonlinear form and directly calculating the conditional mean and variance is not possible. For this reason, the bootstrap procedure is hinged on Monte Carlo simulation of the bootstrap samples. Let $T(X^*)$ denote a statistic computed using the bootstrap sample X^*. To obtain the conditional distribution of $T_b(X^*)$, one can simply draw many (say B) independent copies of the bootstrap sample, and compute $T_b(X^*)$ for each bootstrap sample b. Then $T_1, T_2, ..., T_B$ form an i.i.d. sample from the conditional distribution of $T_b(X^*)$ and this sample can be used to estimate whatever feature of the conditional distribution.

3 Bootstrap Standard Error

I start with the third purpose of bootstrap: to obtain standard error for an estimator, i.e., to use $\sqrt{Var^*(\hat{\theta}_n^*)}$ as an estimator for the standard deviation of an estimator $\hat{\theta}_n$, where $\hat{\theta}_n^*$ is estimator computed using the bootstrap sample and Var^* is the variance taken conditional on the original sample of data. The $\sqrt{Var^*(\hat{\theta}_n^*)}$ is called the (ideal) bootstrap standard error of $\hat{\theta}_n$. In practice, one simulate B bootstrap samples, compute $\{\hat{\theta}_{n,b}^*\}_{b=1}^B$ and use the simulated bootstrap standard error: $\sqrt{(B-1)^{-1}\sum_{b=1}^B(\hat{\theta}_{n,b}^* - \bar{\hat{\theta}}_{n,B}^*)^2}$, where $\bar{\hat{\theta}}_{n,B}^* = B^{-1}\sum_{b=1}^B \hat{\theta}_{n,b}^*$.

Bootstrap standard errors are used as an alternative of the usual asymptotic standard error, the latter being $n^{-1/2}$ times a consistent estimator of the asymptotic standard deviation (squared root of the asymptotic variance) of an $\sqrt{n}$ - asymptotically normal estimator $\hat{\theta}_n$ (i.e. $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, \sigma^2)$ for some $\sigma > 0$). If $\hat{\theta}_n$ is not asymptotically normal, then the use of standard error itself (to build t- or normal-based confidence interval or to conduct t- or normal-based tests) is not appropriate, in which case, it is useless to bootstrap standard errors.

In practice, there are a couple of reasons that bootstrap standard errors are used instead of asymptotic standard errors are as follows. The general feeling is that they work well in the context that they are appropriate (discussed above).

1. Though it is not the case for any of the extremum estimators we have discussed in this course, there are $\sqrt{n}$ - asymptotically normal estimators, whose asymptotic variance involves nonparametric functions. For example, the sample q quantile as an estimator of the population q quantile, τ_q, of a random variable X has an asymptotic variance $1/(q(1-q) * f_X(\tau_q))$, where $f_X(\cdot)$ is the density function of X. Estimators of nonparametric functions usually have slow convergence rate and are sensitive to bandwidth choices. On the other hand, bootstrap standard errors are straightforward to compute and avoids making bandwidth choices.

2. As we have learned in previous lectures, whether a model is correctly specified, or has homoskedastic error (in the nonlinear or linear regression case) affects the shape of the asymptotic variance of extreme estimators. Misspecification and/or heteroskedasticity typically

> complicates the formula greatly. Although closed form formulae can be derived in principle, they are a pain to derive and a even bigger pain to translate into computer code.[1] The "usual" standard errors reported in statistics packages like STATA are therefore often based on the correct-specification and homoskedasticity version of the asymptotic variance formula. Clearly they are not valid if there is misspecification or homoskedasticity. On the other hand the bootstrap standard errors are straighforward to compute, robust to misspecification or heteroskedasticity, and much less susceptible to algebraic or coding errors.

In contrast to the extensive use of bootstrap standard error in applied works and the detailed exploration of the properties of bootstrap tests, the theoretical work on bootstrap standard error is rather incomplete. We shall describe what's available and try to explain why more is not available subsequently. But first, let's define what we mean by the validity of bootstrap standard error. The bootstrap standard error is used to conduct t-tests and construct confidence intervals. Both tasks need

$$\frac{\hat{\theta}_n - \theta_0}{\sqrt{Var^*(\hat{\theta}^*_n)}} \to_d N(0,1). \tag{1}$$

Presumably one already have established $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, \sigma^2)$ for some $\sigma > 0$. Then we only need

$$nVar^*(\hat{\theta}^*_n) \to_p \sigma^2. \tag{2}$$

This condition defines the validity of bootstrap standard error. If the simulated version of the bootstrap standard error is used, this condition should have $Var^*(\hat{\theta}^*_n)$ changed into the simulated variance of the bootstrap estimators.

3.1 Location Model

Consider the location model above again. We would like to use $Var^*(\bar{X}^*_n)$ to approximate $Var(\bar{X})$. Given $\{X_1, ..., X_n\}$, we know that each bootstrap draw X^*_i can take n possible values: $\{X_1, ..., X_n\}$, each with probability $1/n$ and the bootstrap draws are independent from each other. Therefore, we can calculate:

$$\begin{aligned} Var^*(\bar{X}^*_n) &= Var^*(X^*_1)/n = n^{-1}(E^*[(X^*_1)^2] - [E^*(X^*_1)]^2) \\ &= n^{-1}\left(n^{-1}\sum_{i=1}^n X_i^2 - \left(n^{-1}\sum_{i=1}^n X_i\right)^2\right) = n^{-2}\sum_{i=1}^n (X_i - \bar{X}_n)^2. \end{aligned} \tag{3}$$

Thus, the bootstrap standard error is $\hat{\sigma}/\sqrt{n}$ where $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. It turns out that the bootstrap standard error is identical to the standard error computed from asymptotic approx-

[1] Even if one is patient, one might not be careful enough to ensure the derivation and the coding are error free.

imation. Its validity can be verified by an application of the law of large numbers.

3.2 Smooth Functions of the Location

Suppose that the parameter of interest is instead $\theta_0 = g(\mu)$ where $g(\cdot)$ is twice continuously differentiable and has bounded first and second derivatives. Let $\hat{\theta}_n = g(\bar{X}_n)$ and $\hat{\theta}_n^* = g(\bar{X}_n^*)$. Then we have

$$\begin{aligned} |E^* g(\bar{X}_n^*) - g(\bar{X}_n)| &= |E^*(g'(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n)) + (1/2)E^*(g^{''}(\tilde{X}_n)(\bar{X}_n^* - \bar{X}_n)^2)| \\ &\leq \left|\sup_x g^{''}(x)\right| E^*(\bar{X}_n^* - \bar{X}_n)^2 \\ &= \left|\sup_x g^{''}(x)\right| n^{-1}\hat{\sigma}^2 \\ &= O_p(n^{-1}), \end{aligned} \tag{4}$$

where the first equality holds by a second order Taylor expansion, $\tilde{X}$ is some value lying in between $\bar{X}_n^*$ and $\bar{X}_n$, the inequality holds because $E(\bar{X}_n^* - \bar{X}_n) = 0$, the second equality holds by (3) and the last equality holds by standard arguments for sample variances and the boundness assumption on the second derivative of g. Also,

$$\begin{aligned} E^*(g(\bar{X}_n^*) - g(\bar{X}_n))^2 &= E^*[(g'(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n)) + (1/2)(g''(\tilde{X}_n)(\bar{X}_n^* - \bar{X}_n)^2)]^2 \\ &= E^*(g'(\bar{X}_n)^2(\bar{X}_n^* - \bar{X}_n)^2) + E^*(g'(\bar{X}_n)g''(\tilde{X}_n)(\bar{X}_n^* - \bar{X}_n)^3) \\ &\quad + E^*((g''(\tilde{X}_n)^2(\bar{X}_n^* - \bar{X}_n)^4)/4 \\ &= g'(\bar{X}_n)^2 n^{-1}\hat{\sigma}^2 + O_p(n^{-2}), \end{aligned} \tag{5}$$

where the $O_p(n^{-2})$ holds if X has finite fourth moment.

Therefore, we have

$$nVar^*(\hat{\theta}_n^*) = g'(\bar{X}_n)^2\hat{\sigma}^2 + O_p(n^{-1}) \to_p g'(\mu)^2\sigma^2. \tag{6}$$

The limit is exactly the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

We sum up this subsection by the following proposition:

Proposition 10.1 Assume that $g(\cdot)$ is twice continuously differentiable and has bounded first and second derivatives. Also assume that X has finite fourth moment. Then the bootstrap standard error for $\hat{\theta}_n = g(\bar{X}_n)$ is valid in the sense of (2).

3.3 Linear Regression

Now consider a more realistic econometrics model: the linear regression model:

$$Y = X'\beta + \varepsilon, \quad E[\varepsilon|X_i] = 0. \tag{7}$$

The OLS estimator is $\hat{\beta}_n = \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \sum_{i=1}^n X_i Y_i$. Assume that X and ε have finite second moment, and it is easy to show that

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d N(0, V), \tag{8}$$

where $V = (E(X_i X_i'))^{-1} E(X_i X_i' \sigma^2(X_i))(E(X_i X_i'))^{-1}$ and $\sigma^2(X_i) = E(\varepsilon_i^2|X_i)$.

The OLS estimator $\hat{\beta}_n$ is a function of location estimators: $\hat{\mu}_{XX} := n^{-1}\sum_{i=1}^n (X_i X_i')$ and $\hat{\mu}_{XY} := n^{-1}\sum_{i=1}^n (X_i Y_i')$:

$$\hat{\beta}_n = g(\hat{\mu}_{XX}, \hat{\mu}_{XY}),$$

where $g(\mu_{XX}, \mu_{XY}) = \mu_{XX}^{-1}\mu_{XY}$. However, this g function, unlike the ones in the previous subsection, does not have bounded first or second derivative. Thus, the calculation in the previous section does not apply.

In class, we discussed modifications to the nonparametric i.i.d. bootstrap: the residual bootstrap and the wild bootstrap, both of which avoid resampling X_i. Thus in the bootstrap statistics, the $\hat{\mu}_{XX}$ part stays the same as in the original sample statistic and only the $\hat{\mu}_{XY}$ part changes to a bootstrap version. That way, the unbounded first derivative of g with respect to μ_{XX} will not be a problem when calculating the conditional variance of the bootstrap statistic.

As we discussed in class, the convenience in calculating conditional variance of the bootstrap statistic provided by the residual bootstrap and the wild bootstrap does not extend to even nonlinear regression models, let alone more general nonlinear models. Now we discuss another possible way to show bootstrap consistency, which has caveats, but extends more easily to nonlinear models. The approach makes use of the following Theorem from Billingsley (1999):

Theorem. (3.5 of Billingsley (1999)). Let X_n be a sequence of random variables such that $X_n \to_d X$ for some other random variable X. If $\lim_{\alpha\to\infty}\sup_n E[1\{|X_n| \geq \alpha\}|X_n|] = 0$, then $E[|X|] < \infty$ and $EX_n \to EX$.

This suggests that we can show (2) by showing (i) $\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n) \to_d N(0, V)$ conditional on almost all paths of the original sample and (ii) $\lim_{\alpha\to\infty}\sup_n E^*[1\{\|\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n)\| \geq \alpha\}\|\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n)\|^2] = 0$ conditional on almost all paths of the original sample. The first is called the bootstrap distribution consistency and the second is the uniform integrability condition. The uniform integrability condition is difficult to establish. In fact, to the best of my knowledge, there has been no paper that was able to establish the uniform integrability condition for the bootstrap estimator

$\hat{\beta}_n^* = \left(\sum_{i=1}^n X_i^* X_i^{*,\prime}\right)^{-1} \sum_{i=1}^n X_i^* Y_i^*$, let alone the bootstrap versions of more general extremum estimators. The difficulty lies in the inverse of $\left(\sum_{i=1}^n X_i^* X_i^{*,\prime}\right)$ part. The matrix cannot be bounded away from singularity. In fact, it is singular with positive (though very very small) probability causing $\hat{\beta}_n^*$ to be indeterminant. Instead of of getting into details of that, we pretend that the uniform integrability condition is already established.

Consider paths of $\{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), ...\}$ such that $\hat{\mu}_{XX} \to \mu_{XX} := E[X_1 X_1']$, $n^{-1} \sum_{i=1}^n ||(X_i X_i')||^2 < \infty$, $n^{-1} \sum_{i=1}^n (X_i(Y_i - X_i'\hat{\beta}_n))(X_i(Y_i - X_i'\hat{\beta}_n))' \to E[X_1 X_1' \sigma^2(X_1)]$ and $n^{-1} \sum_{i=1}^n ||(X_i(Y_i - X_i'\hat{\beta}_n))||^3 < \infty$. Such paths occur with probability one by the law of large numbers if X_i and Y_i have finite 6th moments. Conditional on such a path, the nonparametric bootstrap sample forms a triangular array of random variables. We can then use the triangular array LLN and CLT to derive the conditional asymptotic distribution. Let $\to_{p^*}$ denote convergence in probability conditional on the paths of original sample described above and similarly for $\to_{d^*}$. First,

$$\hat{\mu}_{XX}^* := n^{-1} \sum_{i=1}^n X_i^* X_i^{*,\prime} \to_{p^*} \mu_{XX},$$

because $E^*[X_i^* X_i^{*,\prime}] = \hat{\mu}_{XX} \to \mu_{XX}$ and $E^*||X_i^* X_i^{*,\prime}||^2 = n^{-1} \sum_{i=1}^n ||(X_i X_i')||^2 < \infty$. Second,

$$\sqrt{n}(\hat{\mu}_{XY}^* - \hat{\mu}_{XX}^* \hat{\beta}_n) = n^{-1/2} \sum_{i=1}^n (X_i^*(Y_i^* - X_i^{*,\prime}\hat{\beta}_n)) \to_{d^*} N(0, E[X_1 X_1' \sigma^2(X_1)]), \tag{9}$$

because $E^*(X_i^*(Y_i^* - X_i^{*,\prime}\hat{\beta}_n)) = n^{-1} \sum_{i=1}^n (X_i(Y_i - X_i'\hat{\beta}_n)) = 0$, $E^*(X_i^*(Y_i^* - X_i^{*,\prime}\hat{\beta}_n))(X_i^*(Y_i^* - X_i^{*,\prime}\hat{\beta}_n))' = n^{-1} \sum_{i=1}^n (X_i(Y_i - X_i'\hat{\beta}_n))(X_i(Y_i - X_i'\hat{\beta}_n))' \to E[X_1 X_1' \sigma^2(X_1)]$, and $E^*||(X_i^*(Y_i^* - X_i^{*,\prime}\hat{\beta}_n))||^3 = n^{-1} \sum_{i=1}^n ||(X_i(Y_i - X_i'\hat{\beta}_n))||^3 < \infty$. Then by the continuous mapping theorem, we have

$$\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n) \to_{d^*} N(0, V). \tag{10}$$

The following proposition sums up this subsection:

Proposition 10.2. (a) Suppose that X_i and Y_i have finite 6th moments.[2] Then the nonparametric i.i.d. bootstrap consistently estimate the asymptotic distribution of the $\sqrt{n}(\hat{\beta}_n - \beta)$.

(b) If in addition, the uniform integrability condition (condition (ii) above) is also satisfied, then the nonparametric i.i.d. bootstrap standard error for the OLS estimator is valid in the sense of (2).

The proof of distributional consistency extends easily to the extremum estimators that we discussed in previous lectures. However, establishing uniform integrability is equally hard, if not harder than the linear regression case.

The theoretical difficulty of establishing uniform integrability does not seem to cause much problem in practice.

[2] In fact, it is sufficient to assume finite $4 + \delta$th moment for any $\delta > 0$.

4 Bias Correction

Bootstrap also offers a way to estimate the bias in certain estimators up to some asymptotic order. For any estimator $\hat{\theta}_n$, the bootstrap bias corrected estimator is

$$\hat{\theta}_n^{\dagger} = \hat{\theta}_n - (E^*(\hat{\theta}_n^*) - \hat{\theta}_n) = 2\hat{\theta}_n - E^*(\hat{\theta}_n^*). \quad (11)$$

Consider the problem in Section 3.2 above. The estimator $\hat{\theta}_n = g(\bar{X}_n)$ for the parameter $\theta_0 = g(\mu)$ is not unbiased if g is not linear. We can characterize the bias by doing a Taylor expansion of $g(\bar{X}_n)$ around μ:

$$\begin{aligned} E[g(\bar{X}_n) - g(\mu)] &= E[g'(\mu)(\bar{X}_n - \mu) + (1/2)g''(\mu)(\bar{X}_n - \mu)^2 + (1/6)(g'''(\tilde{\mu}))(\bar{X}_n - \mu)^3] \\ &= 0 + n^{-1}g''(\mu)\sigma^2/2 + (1/6)E[(g'''(\tilde{\mu}))(\bar{X}_n - \mu)^3] \\ &= 0 + n^{-1}g''(\mu)\sigma^2/2 + o(n^{-1}), \end{aligned} \quad (12)$$

where the last inequality holds if $g'''(\cdot)$ is a bounded function and $E(|X|^3) < \infty$. This shows that the bias in $g(\bar{X}_n)$ is of the order n^{-1}, i.e.

$$nBias(g(\bar{X}_n)) := nE[g(\bar{X}_n) - g(\mu)] \to g''(\mu)\sigma^2/2. \quad (13)$$

If $g''(\mu)$ is large and/or sample size is small, the bias can be substantial. One way to correct the bias is to estimate $g''(\mu)\sigma^2$ by $g''(\bar{X}_n)\hat{\sigma}^2$ and remove n^{-1} times that from the original estimator. Bootstrap provides an alternative to this.

As defined above, the bootstrap use $n\widehat{Bias}^{BT} := nE^*(g(\bar{X}_n^*) - g(\bar{X}_n))$ to estimate $g''(\mu)\sigma^2/2$.

$$\begin{aligned} \widehat{Bias}^{BT} &= E^*(g'(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n)) + (1/2)E^*(g''(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n)^2) + (1/6)g'''(\tilde{X}_n)(\bar{X}_n^* - \bar{X}_n)^3 \\ &= (1/2)g''(\bar{X}_n)E^*(\bar{X}_n^* - \bar{X}_n)^2 + o_p(n^{-1}) \\ &= (1/2)n^{-1}g''(\bar{X}_n)\hat{\sigma}_n^2 + o_p(n^{-1}) \end{aligned} \quad (14)$$

This shows the bootstrap consistency for bias correction,i.e. $n\widehat{Bias}^{BT} \to_p g''(\mu)\sigma^2/2$.

Bias correction usually increases the variance because the bias is estimated. Therefore it should not be used indiscriminately.

Proving bootstrap consistency for bias correction for OLS estimators or other extremum estimators involves the uniform integrability issue as well and thus has the same difficulty as proving bootstrap standard error consistency.

5 Bootstrap Critical Values

The most well-studied use of bootstrap is to obtain critical values for confidence intervals and tests. The reason perhaps is that it is in this context that bootstrap can be theoretically demonstrated to "improve" upon asymptotic approximation. This improvement is often called the "higher-order refinement" of bootstrap approximation. In comparison, bootstrap standard errors are not known to have a similar "refinement" property over asymptotic approximation standard errors, which might be the reason that some bootstrap gurus do not recommend using them.

5.1 Bootstrap confidence interval

The simplest bootstrap confidence interval is the Bootstrap percentile confidence interval of Efron and Tibshirani (1993). For any estimator $\hat{\theta}_n$ of θ. Let $\hat{\theta}_n^*$ be the bootstrap version of $\hat{\theta}_n$. Let q_α^* be the α conditional quantile of $\hat{\theta}_n^*$. Then a Bootstrap percentile confidence interval of nominal coverage probability α is simply

$$CI_{BP} = [q^*_{\alpha_1}, q^*_{1-\alpha_2}]$$

where $\alpha_1 + \alpha_2 = \alpha$. If $\alpha_1 = \alpha_2$, the CI is an equal-tailed CI. If $\alpha_1 = 0$ or $\alpha_2 = 0$, the CI is a one-sided CI. If α_1 and α_2 are chosen to satisfy $\hat{\theta}_n - q^*_{\alpha_1} = q^*_{1-\alpha_2} - \hat{\theta}_n$, then the CI is a symmetric two-sided CI.

The validity of CI_{BP} can be shown by observing that

$$\begin{aligned}\Pr(\theta \in CI_{BP}) &= \Pr(\sqrt{n}(q^*_{\alpha_1} - \hat{\theta}_n) \leq \sqrt{n}(\theta - \hat{\theta}_n) \leq \sqrt{n}(q^*_{1-\alpha_2} - \hat{\theta}_n)) \\ &= \Pr(Q^*_{\alpha_1} \leq \sqrt{n}(\theta - \hat{\theta}_n) \leq Q^*_{1-\alpha_2}), \end{aligned} \tag{15}$$

where Q^*_α is the α conditional quantile of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$. This probability converges to α if we can show that

(i) $\sqrt{n}(\hat{\theta}_n - \theta) \to_d G$ for some symmetric distribution G,

(ii) $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \to_{d^*} G$ conditional on almost all paths of the original sample,

(iii) G has continuous and strictly increasing c.d.f.

Typically G is a normal distribution with an unknown asymptotic variance.

Notice that CI_{BP} is effectively using the quantile of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ to mimic the quantiles of $\sqrt{n}(\theta - \hat{\theta}_n)$. If the exact distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ is skewed, then typically, $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ has similar skewness with $\sqrt{n}(\hat{\theta}_n - \theta)$, but clearly the opposite skewness as $\sqrt{n}(\theta - \hat{\theta}_n)$. That is, CI_{BP} gets the direction of the skewness exactly wrong. Hall (1992) proposes another bootstrap percentile CI, denoted CI_{BP2} that does not get the direction of the skewness wrong. This CI is defined by the inequality

$$Q^*_{\alpha_1} \leq \sqrt{n}(\hat{\theta}_n - \theta) \leq Q^*_{1-\alpha_2}.$$

Thus $CI_{BP2} = [\hat{\theta}_n - Q^*_{1-\alpha_2}/\sqrt{n}, \hat{\theta}_n - Q^*_{\alpha_1}/\sqrt{n}]$. If $\alpha_1 = \alpha_2$, the CI is an equal-tailed CI. If $\alpha_1 = 0$ or $\alpha_2 = 0$, the CI is a one-sided CI. If α_1 and α_2 are chosen to satisfy $-Q^*_{\alpha_1} = Q^*_{1-\alpha_2}$, then the CI is a symmetric two-sided CI.

Both CI_{BP} and CI_{BP2} are **consistent**, in the sense that the coverage probabilities converge to the nominal level. However, neither converges faster than confidence intervals using critical values based on asymptotic approximation (e.g. using the standard normal critical value times an estimator of asymptotic variance in place of Q^*_α when G is known to be a normal with consistently estimable variance). In other words, they **do not offer high-order refinement**.

The reason that they do not offer high-order refinement is that the difference between the conditional distribution of $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n)$ and the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$: $\Pr\left(\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n) \leq x|\mathbb{W}\right) - \Pr(\sqrt{n}(\hat{\theta}_n - \theta) \leq x)$ converges to zero at the same rate as $\Phi(x/\hat{\sigma}_n) - \Pr(\sqrt{n}(\hat{\theta}_n - \theta) \leq x)$, where Φ is the standard normal c.d.f., and $\hat{\sigma}_n$ is a consistent estimator of the asymptotic standard deviation of $\sqrt{n}(\hat{\theta}_n - \theta)$.

But it can be shown under appropriate conditions that $\Pr\left(\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n)/\hat{\sigma}^*_n \leq x|\mathbb{W}\right) - \Pr(\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_n \leq x)$ converges to zero at a faster rate than $\Phi(x) - \Pr(\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_n \leq x)$, where $\hat{\sigma}^*_n$ is $\hat{\sigma}_n$ computed using a bootstrap sample. This suggests that confidence intervals that achieves asymptotic refinement should replace Q^*_α in CI_{BP2} by $z^*_\alpha \hat{\sigma}_n$, where z^*_α is the α quantile of $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n)/\hat{\sigma}^*_n$. And these are the CI's described in lecture 1 of the Andrews notes sent via email.

The CI's that achieve asymptotic refinement obviously requires more coding and more computation because $\hat{\sigma}_n$ has to be coded up and computed – and it has to be coded up correctly. Otherwise, the supposedly "refined" CI's do not achieve refinement, and are often even inconsistent. One easy-to-make mistake when coding $\hat{\sigma}_n$ is to use the asymptotic variance formula for correctly specified model.

5.2 Bootstrap test

The key for bootstrap test is that the bootstrap statistic should mimic the null, not the true distribution of the test statistic. See Andrews notes sent via email.

5.3 Higher-order refinement

Edgeworth expansion. t-test example. GMM modification See Andrews notes sent via email.

But, bootstrapping asymptotically pivotal test statistic does not guarantee high-order refinement. For example, bootstrapping the usual t-statistic for a GMM estimator does not bring high-order refinement because the usual t-statistic uses a standard error formula that assumes correct specification and correct specification does not hold for the bootstrap GMM problem. There are two-solution to this problem, one is Horowitz's solution discussed in Andrews notes, which is to modify the bootstrap GMM problem to force correct specification. The other is to use a standard error formula that is robust to misspecification. (SeoJeong Lee, 2012). Both solutions can achieve high-order refinement.

6 Bootstrap Failure

In general, bootstrap has to be used properly to achieve desired performance. For example,

1. To use bootstrap for standard error, the estimator under consideration must be asymptotically normal. Otherwise, the use of standard error itself is misguided.

2. To use bootstrap critical value in hypothesis tests, the statistic to be computed using the bootstrap sample typically is not exactly the same as the test statistic. The test statistic has to be modified to ensure that the conditional distribution of the bootstrap statistic mimic the null distribution of the test statistic, but not the true distribution of it when the null is not true.

3. To achieve higher-order refinement, the test statistic whose distribution we would like to approximate using bootstrap needs to be asymptotically pivotal.

If the rules are not followed, then bootstrap fails to deliver the desired properties. Thus, "bootstrap failure" is in fact the norm rather than the exceptions. That said, in the econometrics literature, "bootstrap failure" refers to some more specific situations.

The first situation is when the bootstrap standard error is not consistent in the sense of equation (2) even though the estimator under consideration is asymptotically normal. One notable example of such estimator is the nearest neighbor matching estimator. Abadie and Imbens (2008) demonstrates that the bootstrap estimated variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ can either be too large or too small than the asymptotic variance.

The second situation is when the standard bootstrap t-statistic $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$ does not converge to the same distribution as the t-statistic $\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_n$. As a result, the bootstrap percentile CI's do not have the correct asymptotic coverage probability, and t-test using critical values that are quantiles of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$ do not have correct asymptotic null-rejection probability.

One notable example of the second situation was given in Andrews (1997, "A Simple Counter Example to the Bootstrap"). The example considered therein is a location model with nonnegativity constraing. The parameter of interest is the location of X: $\mu = E(X)$. But there is prior information that $\mu \geq 0$. Thus, a reasonable estimator that takes advantage of the prior information is

$$\hat{\mu}_n = [\bar{X}_n]_+ := \max\{0, \bar{X}_n\}.$$

For simplicity, assume that $Var(X) = 1$ is known. Suppose that the true value of μ is on the boundary of the parameter space: $\mu = 0$. For this estimator, the t-statistic has the following asymptotic distribution:

$$\sqrt{n}(\hat{\mu}_n - \mu) = \sqrt{n}([\bar{X}_n]_+ - \mu) \to_d [Z]_+,$$

where $Z \sim N(0,1)$. The bootstrap t-statistic is $\sqrt{n}(\hat{\mu}_n^* - \hat{\mu}) = \sqrt{n}([\bar{X}_n^*]_+ - [\bar{X}_n]_+)$. It does not converge to $[Z]_+$ in distribution conditional on almost all paths $X_1, X_2, ...$ To see why, consider the conditional probability for any $c > 0$,

$$\begin{aligned}\Pr\left(\sqrt{n}(\hat{\mu}_n^* - \hat{\mu}) \leq x | \sqrt{n}\bar{X}_n > c\right) &= \Pr\left(\max\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n), -\sqrt{n}\bar{X}_n\} \leq x | \sqrt{n}\bar{X}_n > c\right) \\ &\geq \Pr\left(\max\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n), -c\} \leq x | \sqrt{n}\bar{X}_n > c\right) \\ &\to \Pr(\max\{Z, -c\} \leq x) \\ &\geq \Pr([Z]_+ \leq x), \end{aligned} \tag{16}$$

and the last inequality is strict if $-c < x < 0$.

This simple example generalizes to extremum estimators when the true parameter lies on the boundary of its parameter space. In thoses cases, the conditional distribution of the bootstrap t-statistic also does not approximate the distribution of the t-statistic, and the bootstrap percentile quantile does not have correct asymptotic coverage probability.