

Lecture Note 6: Bootstrap

One reason why GMM theory is so useful is that it provides a general asymptotic distribution theory for a wide range of estimators. Using the GMM results, we can then construct variance estimates, standard errors, confidence intervals, and hypothesis tests that are asymptotically valid.

Sometimes, it is difficult to calculate variance estimates, confidence intervals, etc., even for relatively straightforward estimators.

Example 1: Nonparametric Bootstrap for Sample Median

Suppose that x_1, \dots, x_n are IID with CDF F and density f . Let

$$\theta = \text{Med}(x_i).$$

Natural estimator: the sample median $\hat{\theta} = \widehat{\text{Med}}(x_1, \dots, x_n)$, which solves

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^n |x_i - \theta|.$$

Assume the density of x at the median, $f(\theta)$, satisfies $0 < f(\theta) < \infty$. Then it can be shown that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V),$$

where

$$V = \frac{1}{4f(\theta)^2}.$$

Problem: how to estimate V ?

One solution is to use a nonparametric density estimator $\hat{f}(\theta)$, and set

$$\hat{V} = \frac{1}{4\hat{f}(\theta)^2}.$$

This requires choosing a bandwidth or smoothing parameter.

An alternative is to construct a bootstrap approximation:

For $b = 1, \dots, B$ (where B is very large), we construct a “bootstrap sample” x_1^b, \dots, x_n^b by drawing with replacement from the observed values $\{x_1, \dots, x_n\}$. Then we construct

$$\hat{\theta}^b = \widehat{\text{Med}}(x_1^b, \dots, x_n^b).$$

Let

$$R^b = \sqrt{n}(\hat{\theta}^b - \hat{\theta}).$$

This leads to a sample R^1, \dots, R^B .

We can then estimate the variance V by

$$\hat{V} = \widehat{Var}(R^1, \dots, R^B).$$

We could then calculate standard errors and construct a confidence interval for θ in the usual way. Alternatively, we can construct a confidence in the following manner:

Using the bootstrap samples, calculate the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of R^b . Call these $\hat{Q}_{\alpha/2}$ and $\hat{Q}_{1-\alpha/2}$. Construct the confidence interval as

$$CI = \{\theta : \hat{Q}_{\alpha/2} \leq \sqrt{n}(\hat{\theta} - \theta) \leq \hat{Q}_{1-\alpha/2}\}.$$

Rearranging, we get

$$CI = \left\{ \theta : \hat{\theta} - \frac{\hat{Q}_{1-\alpha/2}}{\sqrt{n}} \leq \theta \leq \hat{\theta} - \frac{\hat{Q}_{\alpha/2}}{\sqrt{n}} \right\}.$$

It can be shown that

$$\hat{V} \xrightarrow{p} V,$$

and

$$Pr(\theta \in CI) \rightarrow 1 - \alpha \quad \text{with probability 1.}$$

Example 2: Parametric Bootstrap for MLE

Suppose x_i are IID $f(x|\gamma)$, for $\gamma \in \Gamma$, and we are interested in

$$\theta \equiv h(\gamma).$$

For example, θ might be the first element of γ , or some other function of the parameters.

Let $\hat{\gamma}$ be the MLE of γ , and let

$$\hat{\theta} = h(\hat{\gamma}).$$

If the parametric model satisfies regularity conditions,

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, I_\gamma^{-1}),$$

where

$$I_\gamma = E[\nabla_\gamma \log f(x|\gamma) \nabla_\gamma \log f(x|\gamma)'].$$

If $h(\cdot)$ is continuously differentiable, then by the Delta method,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, H(\gamma)'I_{\gamma}^{-1}H(\gamma)),$$

where $H(\gamma) = \nabla_{\gamma}h(\gamma)$.

Parametric bootstrap: for $b = 1, \dots, B$, draw

$$x_1^b, \dots, x_n^b \stackrel{\text{iid}}{\sim} f(x|\hat{\gamma}).$$

Then form

$$\hat{\gamma}^b = \arg \max_{\gamma} \frac{1}{n} \sum_{i=1}^n \log f(x_i^b|\gamma),$$

$$\hat{\theta}^b = h(\hat{\gamma}^b),$$

$$R^b = \sqrt{n}(\hat{\theta}^b - \hat{\theta}).$$

Use the draws R^1, \dots, R^B as before to form variance estimates and confidence intervals.

The same basic approach can be used for many estimators, and in some cases appears to give better approximations than the standard asymptotic methods.

Compared to analyzing the asymptotic distribution of $\hat{\theta}$ and then figuring out some way to consistently estimate the variance, this seems very convenient provided you have a computer to do the bootstrap sampling. Efron coined the name “bootstrap” to suggest “pulling oneself up by one’s bootstraps.”¹

How and when does the bootstrap work? Here is a general framework for understanding the bootstrap.

General Theory

Let $X_n = (x_1, \dots, x_n)$, an IID random sample from a distribution F .

We assume that $F \in \mathcal{F}$, some set of possible distributions.

Example 1: \mathcal{F} is the set of all continuous distributions on the line with nonzero, finite density at the median.

Example 2: $\mathcal{F} = \{F(\cdot|\gamma), \gamma \in \Gamma\}$.

¹According to wikipedia, the phrase “pulling oneself up by one’s bootstraps” derives from the stories told about the Baron von Munchausen. In one story, he was said to lift himself out of a swamp by pulling his own hair. Later, the story involved him pulling himself out by pulling up on his bootstraps.

We are interested in a parameter $\theta(F)$.

Example 1: $\theta(F) = \text{Med}(F)$.

Example 2: $\theta(F) = h(\gamma)$.

Root: a function $R_n(X_n, \theta(F))$, usually real-valued.

Examples:

$$R_n(X_n, \theta(F)) = \sqrt{n}(\hat{\theta}_n - \theta(F)).$$
$$R_n(X_n, \theta(F)) = \frac{\sqrt{n}(\hat{\theta}_n - \theta(F))}{s_n},$$

where s_n is an estimator of a standard deviation. Note that $\hat{\theta}_n$ and s_n are both functions of X_n .

Let $J_n(F)$ be the distribution of $R_n(X_n, \theta(F))$ under F . We also denote the CDF of the root as

$$J_n(t, F) = \text{Pr}_F(R_n(X_n, \theta(F)) \leq t).$$

(The F subscript indicate that the probability is calculated under the distribution F .)

The problem is to figure out the distribution $J_n(F)$.

Pivotal Method:

In some cases, a certain choice for the root results in $J_n(F)$ not depending on F . In this case, the root is called a pivot.

Example: suppose x_1, \dots, x_n are IID $N(\mu, \sigma^2)$. Let

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2.$$

Then it is well known that

$$R_n(X_n, F) = \frac{\hat{\mu} - \mu}{s_n} \sim t_{n-1},$$

the t distribution with $n - 1$ degrees of freedom.

So the distribution of $R_n(X_n, \theta(F))$ does not depend on μ or σ .

Then we can form an exact confidence interval by using the quantiles of the t distribution.

Another example: Suppose that x_i are IID with a continuous distribution F , and we are interested in the entire distribution function. So

$$\theta(F) = F.$$

Let $\hat{F}(\cdot)$ denote the empirical CDF. This is the CDF of the empirical distribution which puts

probability $1/n$ on each observed value of x_i . It can be written as

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^n 1(x_i \leq t).$$

Let

$$R_n(X_n, F) = \sqrt{n} \sup_x |F(x) - \hat{F}(x)|.$$

This is sometimes called the Kolmogorov-Smirnov statistic. It turns out that the distribution of R_n does not depend on F , and its quantiles have been tabulated. We can then test the null hypothesis $H_0 : F = F_0$ by calculating the Kolmogorov-Smirnov statistic and comparing it to the tables.

Asymptotically Pivotal Roots

Exact pivots are unusual, but sometimes we can construct a root such that

$$J_n(t, F) \rightarrow J(t), \quad \text{as } n \rightarrow \infty,$$

at every continuity point of $J(t)$. In other words, $R_n(X_n, F)$ converges in distribution to a limit that does not depend on F .

Example: suppose that x_i are IID F , with bounded mean and variance, and let $\theta(F) = E[x_i]$. Let \bar{x}_n denote the sample mean and let s_n denote the sample standard deviation. The root

$$R_n(X_n, \theta(F)) = \frac{\sqrt{n}(\bar{x}_n - \theta(F))}{s_n} \xrightarrow{d} N(0, 1).$$

Asymptotic Approximation

More generally, $R_n(X_n, \theta(F))$ will converge to a limit distribution which does depend on F . For example, in the sample median case, the root

$$R_n(X_n, \theta(F)) = \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{4f(\theta)^2}\right).$$

Bootstrap Approximation

In this notation, the bootstrap can be described very simply. Let \hat{F} be some estimator of F . The bootstrap approximates

$$J_n(F)$$

by

$$J_n(\hat{F}).$$

In the nonparametric bootstrap, we take \hat{F} to be the empirical distribution of the x_i 's.

In the parametric bootstrap, we take \hat{F} to be $F(\cdot|\hat{\gamma})$, the maximum likelihood estimate of the distribution.

To form a confidence interval, we will use quantiles of the bootstrap approximate distribution $J_n(\hat{F})$. Recall that for a CDF G , its β quantile is defined as

$$G^{-1}(\beta) = \inf\{t : G(t) \geq \beta\}.$$

So the $\alpha/2$ quantile of the bootstrap distribution will be denoted

$$J_n^{-1}(\alpha/2, \hat{F}).$$

Our bootstrap confidence interval is

$$CI_B = \{\theta : J_n^{-1}(\alpha/2, \hat{F}) \leq R_n(X_n, \theta) \leq J_n^{-1}(1 - \alpha/2, \hat{F})\}.$$

So the key is to figure out the quantiles of the bootstrap distribution, which we can do (approximately) by sampling from the bootstrap distribution and forming empirical quantiles.

In some cases, it is possible to figure out $J_n(\hat{F})$ exactly, without resorting to “bootstrap sampling.”

For example, suppose that x_i are IID $N(\mu, \sigma^2)$, and we use the root

$$R_n = \bar{x}_n - \mu.$$

Let $\hat{\mu}, \hat{\sigma}^2$ denote the MLE of μ and σ^2 , and consider the parametric bootstrap. Then the distribution of R_n under $\hat{\mu}, \hat{\sigma}$ is normal with mean 0, and variance $\hat{\sigma}^2/n$.

However, in most cases, it is difficult to analytically calculate $J_n(\hat{F})$. Then we can use bootstrap sampling to simulate this distribution. In practice, we want to use a large value of B to get a good approximation for this distribution.

Validity of the Bootstrap

Typically, we have that $J_n(F)$ converges to some limit distribution $J(F)$. We will also typically choose the estimator \hat{F} to converge to F .

For the bootstrap to be asymptotically valid, we want to show that $J_n(\hat{F})$ converges to $J(F)$.

For this to hold, we will need some sort of smoothness in how $J_n(F)$ depends on F .

Some terminology: a sequence of distribution functions $G_n(\cdot)$ converges weakly to a limit CDF $G(\cdot)$ if it converges pointwise, at every continuity point of G . We write $G_n \rightsquigarrow G$.

So weak convergence of CDFs is equivalent to convergence in distribution of random variables $Y_n \sim G_n$.

We want to study the distribution functions $J_n(\cdot, \hat{F})$. This is somewhat complicated because the \hat{F} are random. First, consider a nonrandom sequence F_n for $n = 1, 2, \dots$. Suppose we can show that for a nonrandom sequence F_n converging to F , that $J_n(\cdot, F_n)$ converges weakly to $J(\cdot, F)$. Then

it seems plausible that $J_n(\cdot, \hat{F})$ also converges weakly to $J(\cdot, F)$. Here is one way to make that precise:

Theorem: let C_F be a set of sequences $\{F_n \in \mathcal{F}\}$ which includes the sequence $\{F, F, F, \dots\}$. Suppose that for every sequence $\{F_n\}$ in C_F , $J_n(F_n)$ converges weakly to $J(F)$. Now, suppose that the sequence of estimated distributions \hat{F} falls in C_F with probability one. Then

$$J_n(\hat{F}) \rightsquigarrow J(F) \quad \text{with probability one.}$$

Note that “with probability one” means that for almost every sequence \hat{F} the statement holds. The result is a trivial consequence of the definitions.

To relate this general result to confidence intervals, recall that our bootstrap confidence intervals were formed using quantiles of $J_n(\hat{F})$. The following result is handy:

Lemma: Suppose that CDFs G_n converge weakly to G . Assume that G is continuous and strictly increasing at $y = G^{-1}(\beta)$. Then

$$G_n^{-1}(\beta) \rightarrow G^{-1}(\beta).$$

Proof: see Politis, Romano, and Wolf.

In words: if G_n converges weakly to G , then the β quantile of G_n converges to the β quantile of G provided G is continuous and strictly increasing at that point.

Using this lemma, we get the following Corollary to the Theorem:

Corollary: If $J(\cdot, F)$ is continuous and strictly increasing at its $\alpha/2$ and $1 - \alpha/2$ quantiles,

$$J_n^{-1}(\alpha/2, \hat{F}) \rightarrow J^{-1}(\alpha/2, F) \quad \text{with probability one.}$$

$$J_n^{-1}(1 - \alpha/2, \hat{F}) \rightarrow J^{-1}(1 - \alpha/2, F) \quad \text{with probability one.}$$

Furthermore, letting CI_B denote the bootstrap confidence interval outlined above,

$$Pr_F(\theta(F) \in CI_B) \rightarrow 1 - \alpha \quad \text{with probability one.}$$

Now, in order to use the Theorem and its Corollary to verify that the bootstrap is valid for a particular problem, we need to specify a class C_F that is large enough so that \hat{F} falls in C_F with probability one, but small enough so that $J_n(\cdot, F_n)$ converges weakly to $J(\cdot, F)$ for every sequence F_n in C_F .

Consider the problem of estimating the mean of a distribution F when F is “nonparametric.” Let x_i be IID F , and let

$$\theta(F) = E_F[x_i] = \int x dF(x).$$

Assume that the variance of x_i is finite:

$$\sigma^2(F) = V_F[x_i] = \int x^2 dF(x) - \theta(F)^2 < \infty.$$

Let

$$R_n(X_n, \theta) = \sqrt{n}(\bar{x}_n - \theta).$$

Proposition Let C_F be the set of sequences F_n such that $F_n \rightsquigarrow F$, $\theta(F_n) \rightarrow \theta(F)$ and $\sigma^2(F_n) \rightarrow \sigma^2(F)$. Then for every sequence F_n in C_F ,

$$J_n(F_n) \rightsquigarrow J(F),$$

where $J(F)$ is the normal distribution with mean 0 and variance $\sigma^2(F)$.

Furthermore, let \hat{F} be the empirical distribution based on the random sample x_1, \dots, x_n . Then $\hat{F} \in C_F$ with probability one.

Therefore, the nonparametric bootstrap confidence interval is asymptotically valid.

Sketch of Proof of Proposition

To show the first part, consider an arbitrary sequence F_n satisfying the conditions (weak convergence, convergence of mean and variance). For each n , construct random variables

$$z_{n1}, \dots, z_{nn} \stackrel{\text{iid}}{\sim} F_n.$$

Let $\bar{z}_n = \frac{1}{n} \sum_{i=1}^n z_{ni}$. Then $J_n(F_n)$ is the distribution of $\sqrt{n}(\bar{z}_n - \theta(F_n))$. By a version of the central limit theorem (the Lindeberg CLT for arrays), it can be shown that

$$\sqrt{n}(\bar{z}_n - \theta(F_n)) \xrightarrow{d} N(0, \sigma^2(F)).$$

Therefore $J_n(F_n) \rightsquigarrow J(F)$.

To show the second part, the strong law of large numbers implies that $\theta(\hat{F}) = \bar{x}_n$ converges almost surely to $\theta(F)$, and $\sigma^2(\hat{F})$ converges almost surely to $\sigma^2(F)$. By a result called the Glivenko-Cantelli theorem, $\sup_x |\hat{F}(x) - F(x)| \rightarrow 0$ with probability one, implying that $\hat{F} \rightsquigarrow F$ with probability one. \square

So we've shown that the bootstrap for the mean is valid. This result turns out to generalize to objects that are similar to means. For example, suppose that

$$\theta(F) = \int \psi(x) dF(x),$$

for a given $\psi(\cdot)$. Suppose that $\psi(x)$ has finite variance. Then essentially the same arguments show that the nonparametric bootstrap is valid.