Chapter 10

The Bootstrap

10.1 Definition of the Bootstrap

Let F denote a distribution function for the population of observations $(y_i, \boldsymbol{x}_i)$. Let

$$T_n = T_n\left((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n), F\right)$$

be a statistic of interest, for example an estimator $\hat{\theta}$ or a t-statistic $\left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Note that we write T_n as possibly a function of F. For example, the t-statistic is a function of the parameter θ which itself is a function of F.

The exact CDF of T_n when the data are sampled from the distribution F is

$$G_n(u, F) = \Pr(T_n \leq u \mid F)$$

In general, $G_n(u, F)$ depends on F, meaning that G changes as F changes.

Ideally, inference would be based on $G_n(u, F)$. This is generally impossible since F is unknown.

Asymptotic inference is based on approximating $G_n(u, F)$ with $G(u, F) = \lim_{n\to\infty} G_n(u, F)$. When $G(u, F) = G(u)$ does not depend on F, we say that T_n is asymptotically pivotal and use the distribution function $G(u)$ for inferential purposes.

In a seminal contribution, Efron (1979) proposed the bootstrap, which makes a different approximation. The unknown F is replaced by a consistent estimate F_n (one choice is discussed in the next section). Plugged into $G_n(u, F)$ we obtain

$$G_n^*(u) = G_n(u, F_n). \tag{10.1}$$

We call G_n^* the bootstrap distribution. Bootstrap inference is based on $G_n^*(u)$.

Let $(y_i^*, \boldsymbol{x}_i^*)$ denote random variables with the distribution F_n. A random sample from this distribution is called the bootstrap data. The statistic $T_n^* = T_n\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), F_n\right)$ constructed on this sample is a random variable with distribution G_n^*. That is, $\Pr(T_n^* \leq u) = G_n^*(u)$. We call T_n^* the bootstrap statistic. The distribution of T_n^* is identical to that of T_n when the true CDF is F_n rather than F.

The bootstrap distribution is itself random, as it depends on the sample through the estimator F_n.

In the next sections we describe computation of the bootstrap distribution.

10.2 The Empirical Distribution Function

Recall that $F(y, \boldsymbol{x}) = \Pr\left(y_i \leq y, \boldsymbol{x}_i \leq \boldsymbol{x}\right) = \mathbb{E}\left(1\left(y_i \leq y\right) 1\left(\boldsymbol{x}_i \leq \boldsymbol{x}\right)\right)$, where $1(\cdot)$ is the indicator function. This is a population moment. The method of moments estimator is the corresponding

sample moment:

$$F_n(y, \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq y)\, 1(\boldsymbol{x}_i \leq \boldsymbol{x}). \tag{10.2}$$

$F_n(y, \boldsymbol{x})$ is called the empirical distribution function (EDF). F_n is a nonparametric estimate of F. Note that while F may be either discrete or continuous, F_n is by construction a step function.

The EDF is a consistent estimator of the CDF. To see this, note that for any $(y, \boldsymbol{x})$, $1(y_i \leq y)\, 1(\boldsymbol{x}_i \leq \boldsymbol{x})$ is an iid random variable with expectation $F(y, \boldsymbol{x})$. Thus by the WLLN (Theorem 2.6.1), $F_n(y, \boldsymbol{x}) \xrightarrow{p} F(y, \boldsymbol{x})$. Furthermore, by the CLT (Theorem 2.8.1),

$$\sqrt{n}\left(F_n(y, \boldsymbol{x}) - F(y, \boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}\left(0, F(y, \boldsymbol{x})\left(1 - F(y, \boldsymbol{x})\right)\right).$$

To see the effect of sample size on the EDF, in the Figure below, I have plotted the EDF and true CDF for three random samples of size $n = 25$, 50, 100, and 500. The random draws are from the $\mathrm{N}(0, 1)$ distribution. For $n = 25$, the EDF is only a crude approximation to the CDF, but the approximation appears to improve for the large n. In general, as the sample size gets larger, the EDF step function gets uniformly close to the true CDF.

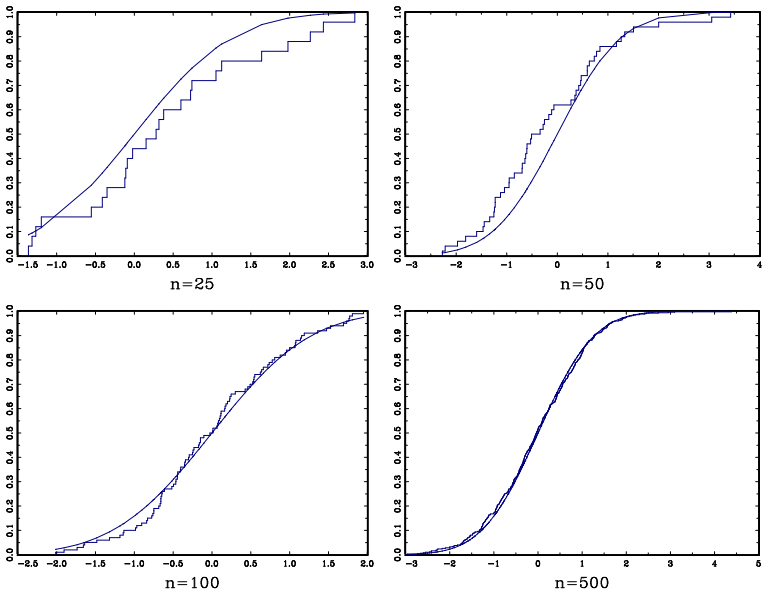

Figure 10.1: Empirical Distribution Functions

The EDF is a valid discrete probability distribution which puts probability mass $1/n$ at each pair $(y_i, \boldsymbol{x}_i)$, $i = 1, ..., n$. Notationally, it is helpful to think of a random pair $(y_i^*, \boldsymbol{x}_i^*)$ with the distribution F_n. That is,

$$\Pr(y_i^* \leq y, \boldsymbol{x}_i^* \leq \boldsymbol{x}) = F_n(y, \boldsymbol{x}).$$

We can easily calculate the moments of functions of $(y_i^*, \boldsymbol{x}_i^*)$:

$$\begin{aligned}
\mathbb{E}h(y_i^*, \boldsymbol{x}_i^*) &= \int h(y, \boldsymbol{x}) dF_n(y, \boldsymbol{x}) \\
&= \sum_{i=1}^{n} h(y_i, \boldsymbol{x}_i) \Pr(y_i^* = y_i, \boldsymbol{x}_i^* = \boldsymbol{x}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} h(y_i, \boldsymbol{x}_i),
\end{aligned}$$

the empirical sample average.

10.3 Nonparametric Bootstrap

The **nonparametric bootstrap** is obtained when the bootstrap distribution (10.1) is defined using the EDF (10.2) as the estimate F_n of F.

Since the EDF F_n is a multinomial (with n support points), in principle the distribution G_n^* could be calculated by direct methods. However, as there are $\binom{2n-1}{n}$ possible samples $\{(y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*)\}$, such a calculation is computationally infeasible. The popular alternative is to use simulation to approximate the distribution. The algorithm is identical to our discussion of Monte Carlo simulation, with the following points of clarification:

- The sample size n used for the simulation is the same as the sample size.
- The random vectors $(y_i^*, \boldsymbol{x}_i^*)$ are drawn randomly from the empirical distribution. This is equivalent to sampling a pair $(y_i, \boldsymbol{x}_i)$ randomly from the sample.

The bootstrap statistic $T_n^* = T_n\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), F_n\right)$ is calculated for each bootstrap sample. This is repeated B times. B is known as the number of bootstrap replications. A theory for the determination of the number of bootstrap replications B has been developed by Andrews and Buchinsky (2000). It is desirable for B to be large, so long as the computational costs are reasonable. $B = 1000$ typically suffices.

When the statistic T_n is a function of F, it is typically through dependence on a parameter. For example, the t-ratio $\left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ depends on θ. As the bootstrap statistic replaces F with F_n, it similarly replaces θ with θ_n, the value of θ implied by F_n. Typically $\theta_n = \hat{\theta}$, the parameter estimate. (When in doubt use $\hat{\theta}$.)

Sampling from the EDF is particularly easy. Since F_n is a discrete probability distribution putting probability mass $1/n$ at each sample point, sampling from the EDF is equivalent to random sampling a pair $(y_i, \boldsymbol{x}_i)$ from the observed data **with replacement**. In consequence, a bootstrap sample $\{(y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*)\}$ will necessarily have some ties and multiple values, which is generally not a problem.

10.4 Bootstrap Estimation of Bias and Variance

The bias of $\hat{\theta}$ is $\tau_n = \mathbb{E}(\hat{\theta} - \theta_0)$. Let $T_n(\theta) = \hat{\theta} - \theta$. Then $\tau_n = \mathbb{E}(T_n(\theta_0))$. The bootstrap counterparts are $\hat{\theta}^* = \hat{\theta}((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*))$ and $T_n^* = \hat{\theta}^* - \theta_n = \hat{\theta}^* - \hat{\theta}$. The bootstrap estimate of τ_n is

$$\tau_n^* = \mathbb{E}(T_n^*).$$

If this is calculated by the simulation described in the previous section, the estimate of τ_n^* is

$$\begin{aligned} \hat{\tau}_n^* &= \frac{1}{B}\sum_{b=1}^{B} T_{nb}^* \\ &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^* - \hat{\theta} \\ &= \overline{\hat{\theta}^*} - \hat{\theta}. \end{aligned}$$

If $\hat{\theta}$ is biased, it might be desirable to construct a biased-corrected estimator (one with reduced bias). Ideally, this would be

$$\tilde{\theta} = \hat{\theta} - \tau_n,$$

but τ_n is unknown. The (estimated) bootstrap biased-corrected estimator is

$$\begin{aligned}\tilde{\theta}^* &= \hat{\theta} - \hat{\tau}_n^* \\ &= \hat{\theta} - (\overline{\hat{\theta}^*} - \hat{\theta}) \\ &= 2\hat{\theta} - \overline{\hat{\theta}^*}.\end{aligned}$$

Note, in particular, that the biased-corrected estimator is *not* $\overline{\hat{\theta}^*}$. Intuitively, the bootstrap makes the following experiment. Suppose that $\hat{\theta}$ is the truth. Then what is the average value of $\hat{\theta}$ calculated from such samples? The answer is $\overline{\hat{\theta}^*}$. If this is lower than $\hat{\theta}$, this suggests that the estimator is *downward-biased*, so a biased-corrected estimator of θ should be *larger* than $\hat{\theta}$, and the best guess is the difference between $\hat{\theta}$ and $\overline{\hat{\theta}^*}$. Similarly if $\overline{\hat{\theta}^*}$ is higher than $\hat{\theta}$, then the estimator is upward-biased and the biased-corrected estimator should be lower than $\hat{\theta}$.

Let $T_n = \hat{\theta}$. The variance of $\hat{\theta}$ is

$$V_n = \mathbb{E}(T_n - \mathbb{E}T_n)^2.$$

Let $T_n^* = \hat{\theta}^*$. It has variance

$$V_n^* = \mathbb{E}(T_n^* - \mathbb{E}T_n^*)^2.$$

The simulation estimate is

$$\hat{V}_n^* = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2.$$

A bootstrap standard error for $\hat{\theta}$ is the square root of the bootstrap estimate of variance, $s^*(\hat{\theta}) = \sqrt{\hat{V}_n^*}$.

While this standard error may be calculated and reported, it is not clear if it is useful. The primary use of asymptotic standard errors is to construct asymptotic confidence intervals, which are based on the asymptotic normal approximation to the t-ratio. However, the use of the bootstrap presumes that such asymptotic approximations might be poor, in which case the normal approximation is suspected. It appears superior to calculate bootstrap confidence intervals, and we turn to this next.

10.5 Percentile Intervals

For a distribution function $G_n(u, F)$, let $q_n(\alpha, F)$ denote its quantile function. This is the function which solves

$$G_n(q_n(\alpha, F), F) = \alpha.$$

[When $G_n(u, F)$ is discrete, $q_n(\alpha, F)$ may be non-unique, but we will ignore such complications.] Let $q_n(\alpha)$ denote the quantile function of the true sampling distribution, and $q_n^*(\alpha) = q_n(\alpha, F_n)$ denote the quantile function of the bootstrap distribution. Note that this function will change depending on the underlying statistic T_n whose distribution is G_n.

Let $T_n = \hat{\theta}$, an estimate of a parameter of interest. In $(1-\alpha)\%$ of samples, $\hat{\theta}$ lies in the region $[q_n(\alpha/2), q_n(1-\alpha/2)]$. This motivates a confidence interval proposed by Efron:

$$C_1 = [q_n^*(\alpha/2), \quad q_n^*(1-\alpha/2)].$$

This is often called the *percentile confidence interval*.

Computationally, the quantile $q_n^*(\alpha)$ is estimated by $\hat{q}_n^*(\alpha)$, the α'th sample quantile of the simulated statistics $\{T_{n1}^*, ..., T_{nB}^*\}$, as discussed in the section on Monte Carlo simulation. The $(1-\alpha)\%$ Efron percentile interval is then $[\hat{q}_n^*(\alpha/2), \quad \hat{q}_n^*(1-\alpha/2)]$.

The interval C_1 is a popular bootstrap confidence interval often used in empirical practice. This is because it is easy to compute, simple to motivate, was popularized by Efron early in the history of the bootstrap, and also has the feature that it is translation invariant. That is, if we define $\phi = f(\theta)$ as the parameter of interest for a monotonically increasing function f, then percentile method applied to this problem will produce the confidence interval $[f(q_n^*(\alpha/2)), \quad f(q_n^*(1-\alpha/2))]$, which is a naturally good property.

However, as we show now, C_1 is in a deep sense very poorly motivated.

It will be useful if we introduce an alternative definition of C_1. Let $T_n(\theta) = \hat{\theta} - \theta$ and let $q_n(\alpha)$ be the quantile function of its distribution. (These are the original quantiles, with θ subtracted.) Then C_1 can alternatively be written as

$$C_1 = [\hat{\theta} + q_n^*(\alpha/2), \quad \hat{\theta} + q_n^*(1-\alpha/2)].$$

This is a bootstrap estimate of the "ideal" confidence interval

$$C_1^0 = [\hat{\theta} + q_n(\alpha/2), \quad \hat{\theta} + q_n(1-\alpha/2)].$$

The latter has coverage probability

$$\begin{aligned} \Pr\left(\theta_0 \in C_1^0\right) &= \Pr\left(\hat{\theta} + q_n(\alpha/2) \le \theta_0 \le \hat{\theta} + q_n(1-\alpha/2)\right) \\ &= \Pr\left(-q_n(1-\alpha/2) \le \hat{\theta} - \theta_0 \le -q_n(\alpha/2)\right) \\ &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0) \end{aligned}$$

which generally is not $1-\alpha$! There is one important exception. If $\hat{\theta}-\theta_0$ has a symmetric distribution about 0, then $G_n(-u, F_0) = 1 - G_n(u, F_0)$, so

$$\begin{aligned} \Pr\left(\theta_0 \in C_1^0\right) &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0) \\ &= (1 - G_n(q_n(\alpha/2), F_0)) - (1 - G_n(q_n(1-\alpha/2), F_0)) \\ &= \left(1 - \frac{\alpha}{2}\right) - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) \\ &= 1 - \alpha \end{aligned}$$

and this idealized confidence interval is accurate. Therefore, C_1^0 and C_1 are designed for the case that $\hat{\theta}$ has a symmetric distribution about θ_0.

When $\hat{\theta}$ does not have a symmetric distribution, C_1 may perform quite poorly.

However, by the translation invariance argument presented above, it also follows that if there exists some monotonically increasing transformation $f(\cdot)$ such that $f(\hat{\theta})$ is symmetrically distributed about $f(\theta_0)$, then the idealized percentile bootstrap method will be accurate.

Based on these arguments, many argue that the percentile interval should not be used unless the sampling distribution is close to unbiased and symmetric.

The problems with the percentile method can be circumvented, at least in principle, by an alternative method.

Let $T_n(\theta) = \hat{\theta} - \theta$. Then

$$\begin{aligned} 1 - \alpha &= \Pr\left(q_n(\alpha/2) \le T_n(\theta_0) \le q_n(1-\alpha/2)\right) \\ &= \Pr\left(\hat{\theta} - q_n(1-\alpha/2) \le \theta_0 \le \hat{\theta} - q_n(\alpha/2)\right), \end{aligned}$$

so an exact $(1-\alpha)\%$ confidence interval for θ_0 would be

$$C_2^0 = [\hat{\theta} - q_n(1-\alpha/2), \quad \hat{\theta} - q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_2 = [\hat{\theta} - q_n^*(1-\alpha/2), \quad \hat{\theta} - q_n^*(\alpha/2)].$$

Notice that generally this is very different from the Efron interval C_1! They coincide in the special case that $G_n^*(u)$ is symmetric about $\hat{\theta}$, but otherwise they differ.

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)$, which are centered at the sample estimate $\hat{\theta}$. These are sorted to yield the quantile estimates $\hat{q}_n^*(.025)$ and $\hat{q}_n^*(.975)$. The 95% confidence interval is then $[\hat{\theta} - \hat{q}_n^*(.975), \quad \hat{\theta} - \hat{q}_n^*(.025)]$.

This confidence interval is discussed in most theoretical treatments of the bootstrap, but is not widely used in practice.

10.6 Percentile-t Equal-Tailed Interval

Suppose we want to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta < \theta_0$ at size α. We would set $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $T_n(\theta_0) < c$, where c would be selected so that

$$\Pr\left(T_n(\theta_0) < c\right) = \alpha.$$

Thus $c = q_n(\alpha)$. Since this is unknown, a bootstrap test replaces $q_n(\alpha)$ with the bootstrap estimate $q_n^*(\alpha)$, and the test rejects if $T_n(\theta_0) < q_n^*(\alpha)$.

Similarly, if the alternative is $\mathbb{H}_1 : \theta > \theta_0$, the bootstrap test rejects if $T_n(\theta_0) > q_n^*(1-\alpha)$.

Computationally, these critical values can be estimated from a bootstrap simulation by sorting the bootstrap t-statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)/s(\hat{\theta}^*)$. Note, and this is important, that the bootstrap test statistic is centered at the estimate $\hat{\theta}$, and the standard error $s(\hat{\theta}^*)$ is calculated on the bootstrap sample. These t-statistics are sorted to find the estimated quantiles $\hat{q}_n^*(\alpha)$ and/or $\hat{q}_n^*(1-\alpha)$.

Let $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Then taking the intersection of two one-sided intervals,

$$\begin{aligned} 1-\alpha &= \Pr\left(q_n(\alpha/2) \le T_n(\theta_0) \le q_n(1-\alpha/2)\right) \\ &= \Pr\left(q_n(\alpha/2) \le \left(\hat{\theta} - \theta_0\right)/s(\hat{\theta}) \le q_n(1-\alpha/2)\right) \\ &= \Pr\left(\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2) \le \theta_0 \le \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)\right), \end{aligned}$$

so an exact $(1-\alpha)$% confidence interval for θ_0 would be

$$C_3^0 = [\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_3 = [\hat{\theta} - s(\hat{\theta})q_n^*(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n^*(\alpha/2)].$$

This is often called a *percentile-t confidence interval.* It is *equal-tailed* or *central* since the probability that θ_0 is below the left endpoint approximately equals the probability that θ_0 is above the right endpoint, each $\alpha/2$.

Computationally, this is based on the critical values from the one-sided hypothesis tests, discussed above.

10.7 Symmetric Percentile-t Intervals

Suppose we want to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ at size α. We would set $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $|T_n(\theta_0)| > c$, where c would be selected so that

$$\Pr\left(|T_n(\theta_0)| > c\right) = \alpha.$$

Note that

$$\begin{aligned}\Pr\left(|T_n(\theta_0)| < c\right) &= \Pr\left(-c < T_n(\theta_0) < c\right) \\ &= G_n(c) - G_n(-c) \\ &\equiv \overline{G}_n(c),\end{aligned}$$

which is a symmetric distribution function. The ideal critical value $c = q_n(\alpha)$ solves the equation

$$\overline{G}_n(q_n(\alpha)) = 1 - \alpha.$$

Equivalently, $q_n(\alpha)$ is the $1-\alpha$ quantile of the distribution of $|T_n(\theta_0)|$.

The bootstrap estimate is $q_n^*(\alpha)$, the $1-\alpha$ quantile of the distribution of $|T_n^*|$, or the number which solves the equation

$$\overline{G}_n^*(q_n^*(\alpha)) = G_n^*(q_n^*(\alpha)) - G_n^*(-q_n^*(\alpha)) = 1 - \alpha.$$

Computationally, $q_n^*(\alpha)$ is estimated from a bootstrap simulation by sorting the bootstrap t-statistics $|T_n^*| = \left|\hat{\theta}^* - \hat{\theta}\right| / s(\hat{\theta}^*)$, and taking the upper α% quantile. The bootstrap test rejects if $|T_n(\theta_0)| > q_n^*(\alpha)$.

Let

$$C_4 = [\hat{\theta} - s(\hat{\theta})q_n^*(\alpha), \quad \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)],$$

where $q_n^*(\alpha)$ is the bootstrap critical value for a two-sided hypothesis test. C_4 is called the symmetric percentile-t interval. It is designed to work well since

$$\begin{aligned}\Pr\left(\theta_0 \in C_4\right) &= \Pr\left(\hat{\theta} - s(\hat{\theta})q_n^*(\alpha) \le \theta_0 \le \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)\right) \\ &= \Pr\left(|T_n(\theta_0)| < q_n^*(\alpha)\right) \\ &\simeq \Pr\left(|T_n(\theta_0)| < q_n(\alpha)\right) \\ &= 1 - \alpha.\end{aligned}$$

If $\boldsymbol{\theta}$ is a vector, then to test $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ at size α, we would use a Wald statistic

$$W_n(\boldsymbol{\theta}) = n\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$$

or some other asymptotically chi-square statistic. Thus here $T_n(\boldsymbol{\theta}) = W_n(\boldsymbol{\theta})$. The ideal test rejects if $W_n \ge q_n(\alpha)$, where $q_n(\alpha)$ is the $(1-\alpha)$% quantile of the distribution of W_n. The bootstrap test rejects if $W_n \ge q_n^*(\alpha)$, where $q_n^*(\alpha)$ is the $(1-\alpha)$% quantile of the distribution of

$$W_n^* = n\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right)' \hat{\boldsymbol{V}}_{\theta}^{*-1}\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right).$$

Computationally, the critical value $q_n^*(\alpha)$ is found as the quantile from simulated values of W_n^*. Note in the simulation that the Wald statistic is a quadratic form in $\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right)$, not $\left(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0\right)$. [This is a typical mistake made by practitioners.]

10.8 Asymptotic Expansions

Let $T_n \in \mathbb{R}$ be a statistic such that

$$T_n \xrightarrow{d} \mathrm{N}(0, \sigma^2). \tag{10.3}$$

In some cases, such as when T_n is a t-ratio, then $\sigma^2 = 1$. In other cases σ^2 is unknown. Equivalently, writing $T_n \sim G_n(u, F)$ then for each u and F

$$\lim_{n\to\infty} G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right),$$

or

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + o(1). \tag{10.4}$$

While (10.4) says that G_n converges to $\Phi\left(\frac{u}{\sigma}\right)$ as $n \to \infty$, it says nothing, however, about the rate of convergence, or the size of the divergence for any particular sample size n. A better asymptotic approximation may be obtained through an *asymptotic expansion.*

The following notation will be helpful. Let a_n be a sequence.

Definition 10.8.1 $a_n = o(1)$ *if* $a_n \to 0$ *as* $n \to \infty$

Definition 10.8.2 $a_n = O(1)$ *if* $|a_n|$ *is uniformly bounded.*

Definition 10.8.3 $a_n = o(n^{-r})$ *if* $n^r |a_n| \to 0$ *as* $n \to \infty$.

Basically, $a_n = O(n^{-r})$ if it declines to zero like n^{-r}.

We say that a function $g(u)$ is *even* if $g(-u) = g(u)$, and a function $h(u)$ is *odd* if $h(-u) = -h(u)$. The derivative of an even function is odd, and vice-versa.

Theorem 10.8.1 *Under regularity conditions and (10.3),*

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + \frac{1}{n^{1/2}} g_1(u, F) + \frac{1}{n} g_2(u, F) + O(n^{-3/2})$$

uniformly over u, where g_1 is an even function of u, and g_2 is an odd function of u. Moreover, g_1 and g_2 are differentiable functions of u and continuous in F relative to the supremum norm on the space of distribution functions.

The expansion in Theorem 10.8.1 is often called an **Edgeworth expansion**.

We can interpret Theorem 10.8.1 as follows. First, $G_n(u, F)$ converges to the normal limit at rate $n^{1/2}$. To a second order of approximation,

$$G_n(u, F) \approx \Phi\left(\frac{u}{\sigma}\right) + n^{-1/2} g_1(u, F).$$

Since the derivative of g_1 is odd, the density function is skewed. To a third order of approximation,

$$G_n(u, F) \approx \Phi\left(\frac{u}{\sigma}\right) + n^{-1/2} g_1(u, F) + n^{-1} g_2(u, F)$$

which adds a symmetric non-normal component to the approximate density (for example, adding leptokurtosis).

[Side Note: When $T_n = \sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$, a standardized sample mean, then

$$\begin{aligned} g_1(u) &= -\frac{1}{6}\kappa_3\left(u^2-1\right)\phi(u) \\ g_2(u) &= -\left(\frac{1}{24}\kappa_4\left(u^3-3u\right)+\frac{1}{72}\kappa_3^2\left(u^5-10u^3+15u\right)\right)\phi(u) \end{aligned}$$

where $\phi(u)$ is the standard normal pdf, and

$$\begin{aligned} \kappa_3 &= \mathbb{E}\left(X-\mu\right)^3/\sigma^3 \\ \kappa_4 &= \mathbb{E}\left(X-\mu\right)^4/\sigma^4 - 3 \end{aligned}$$

the standardized skewness and excess kurtosis of the distribution of X. Note that when $\kappa_3 = 0$ and $\kappa_4 = 0$, then $g_1 = 0$ and $g_2 = 0$, so the second-order Edgeworth expansion corresponds to the normal distribution.]

Francis Edgeworth

Francis Ysidro Edgeworth (1845-1926) of Ireland, founding editor of the *Economic Journal*, was a profound economic and statistical theorist, developing the theories of indifference curves and asymptotic expansions. He also could be viewed as the first econometrician due to his early use of mathematical statistics in the study of economic data.

10.9 One-Sided Tests

Using the expansion of Theorem 10.8.1, we can assess the accuracy of one-sided hypothesis tests and confidence regions based on an asymptotically normal t-ratio T_n. An asymptotic test is based on $\Phi(u)$.

To the second order, the exact distribution is

$$\Pr\left(T_n < u\right) = G_n(u, F_0) = \Phi(u) + \frac{1}{n^{1/2}}g_1(u, F_0) + O(n^{-1})$$

since $\sigma = 1$. The difference is

$$\begin{aligned} \Phi(u) - G_n(u, F_0) &= \frac{1}{n^{1/2}}g_1(u, F_0) + O(n^{-1}) \\ &= O(n^{-1/2}), \end{aligned}$$

so the order of the error is $O(n^{-1/2})$.

A bootstrap test is based on $G_n^*(u)$, which from Theorem 10.8.1 has the expansion

$$G_n^*(u) = G_n(u, F_n) = \Phi(u) + \frac{1}{n^{1/2}}g_1(u, F_n) + O(n^{-1}).$$

Because $\Phi(u)$ appears in both expansions, the difference between the bootstrap distribution and the true distribution is

$$G_n^*(u) - G_n(u, F_0) = \frac{1}{n^{1/2}}\left(g_1(u, F_n) - g_1(u, F_0)\right) + O(n^{-1}).$$

Since F_n converges to F at rate $\sqrt{n}$, and g_1 is continuous with respect to F, the difference $(g_1(u, F_n) - g_1(u, F_0))$ converges to 0 at rate $\sqrt{n}$. Heuristically,

$$\begin{aligned} g_1(u, F_n) - g_1(u, F_0) &\approx \frac{\partial}{\partial F} g_1(u, F_0)\,(F_n - F_0) \\ &= O(n^{-1/2}), \end{aligned}$$

The "derivative" $\frac{\partial}{\partial F} g_1(u, F)$ is only heuristic, as F is a function. We conclude that

$$G_n^*(u) - G_n(u, F_0) = O(n^{-1}),$$

or

$$\Pr\left(T_n^* \le u\right) = \Pr\left(T_n \le u\right) + O(n^{-1}),$$

which is an improved rate of convergence over the asymptotic test (which converged at rate $O(n^{-1/2})$). This rate can be used to show that one-tailed bootstrap inference based on the t-ratio achieves a so-called *asymptotic refinement* – the Type I error of the test converges at a faster rate than an analogous asymptotic test.

10.10 Symmetric Two-Sided Tests

If a random variable y has distribution function $H(u) = \Pr(y \le u)$, then the random variable $|y|$ has distribution function

$$\overline{H}(u) = H(u) - H(-u)$$

since

$$\begin{aligned} \Pr\left(|y| \le u\right) &= \Pr\left(-u \le y \le u\right) \\ &= \Pr\left(y \le u\right) - \Pr\left(y \le -u\right) \\ &= H(u) - H(-u). \end{aligned}$$

For example, if $Z \sim \mathrm{N}(0, 1)$, then $|Z|$ has distribution function

$$\overline{\Phi}(u) = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1.$$

Similarly, if T_n has exact distribution $G_n(u, F)$, then $|T_n|$ has the distribution function

$$\overline{G}_n(u, F) = G_n(u, F) - G_n(-u, F).$$

A two-sided hypothesis test rejects $\mathbb{H}_0$ for large values of $|T_n|$. Since $T_n \xrightarrow{d} Z$, then $|T_n| \xrightarrow{d} |Z| \sim \overline{\Phi}$. Thus asymptotic critical values are taken from the $\overline{\Phi}$ distribution, and exact critical values are taken from the $\overline{G}_n(u, F_0)$ distribution. From Theorem 10.8.1, we can calculate that

$$\begin{aligned} \overline{G}_n(u, F) &= G_n(u, F) - G_n(-u, F) \\ &= \left(\Phi(u) + \frac{1}{n^{1/2}} g_1(u, F) + \frac{1}{n} g_2(u, F)\right) \\ &\quad - \left(\Phi(-u) + \frac{1}{n^{1/2}} g_1(-u, F) + \frac{1}{n} g_2(-u, F)\right) + O(n^{-3/2}) \\ &= \overline{\Phi}(u) + \frac{2}{n} g_2(u, F) + O(n^{-3/2}), \end{aligned} \tag{10.5}$$

where the simplifications are because g_1 is even and g_2 is odd. Hence the difference between the asymptotic distribution and the exact distribution is

$$\overline{\Phi}(u) - \overline{G}_n(u, F_0) = \frac{2}{n} g_2(u, F_0) + O(n^{-3/2}) = O(n^{-1}).$$

The order of the error is $O(n^{-1})$.

Interestingly, the asymptotic two-sided test has a better coverage rate than the asymptotic one-sided test. This is because the first term in the asymptotic expansion, g_1, is an even function, meaning that the errors in the two directions exactly cancel out.

Applying (10.5) to the bootstrap distribution, we find

$$\overline{G}_n^*(u) = \overline{G}_n(u, F_n) = \overline{\Phi}(u) + \frac{2}{n} g_2(u, F_n) + O(n^{-3/2}).$$

Thus the difference between the bootstrap and exact distributions is

$$\begin{aligned} \overline{G}_n^*(u) - \overline{G}_n(u, F_0) &= \frac{2}{n} \left(g_2(u, F_n) - g_2(u, F_0) \right) + O(n^{-3/2}) \\ &= O(n^{-3/2}), \end{aligned}$$

the last equality because F_n converges to F_0 at rate $\sqrt{n}$, and g_2 is continuous in F. Another way of writing this is

$$\Pr\left(|T_n^*| < u\right) = \Pr\left(|T_n| < u\right) + O(n^{-3/2})$$

so the error from using the bootstrap distribution (relative to the true unknown distribution) is $O(n^{-3/2})$. This is in contrast to the use of the asymptotic distribution, whose error is $O(n^{-1})$. Thus a two-sided bootstrap test also achieves an asymptotic refinement, similar to a one-sided test.

A reader might get confused between the two simultaneous effects. Two-sided tests have better rates of convergence than the one-sided tests, and bootstrap tests have better rates of convergence than asymptotic tests.

The analysis shows that there may be a trade-off between one-sided and two-sided tests. Two-sided tests will have more accurate size (Reported Type I error), but one-sided tests might have more power against alternatives of interest. Confidence intervals based on the bootstrap can be asymmetric if based on one-sided tests (equal-tailed intervals) and can therefore be more informative and have smaller length than symmetric intervals. Therefore, the choice between symmetric and equal-tailed confidence intervals is unclear, and needs to be determined on a case-by-case basis.

10.11 Percentile Confidence Intervals

To evaluate the coverage rate of the percentile interval, set $T_n = \sqrt{n}\left(\hat{\theta} - \theta_0\right)$. We know that $T_n \xrightarrow{d} \mathrm{N}(0, V)$, which is not pivotal, as it depends on the unknown V. Theorem 10.8.1 shows that a first-order approximation

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}),$$

where $\sigma = \sqrt{V}$, and for the bootstrap

$$G_n^*(u) = G_n(u, F_n) = \Phi\left(\frac{u}{\hat{\sigma}}\right) + O(n^{-1/2}),$$

where $\hat{\sigma} = V(F_n)$ is the bootstrap estimate of σ. The difference is

$$\begin{aligned} G_n^*(u) - G_n(u, F_0) &= \Phi\left(\frac{u}{\hat{\sigma}}\right) - \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}) \\ &= -\phi\left(\frac{u}{\sigma}\right) \frac{u}{\sigma} \left(\hat{\sigma} - \sigma\right) + O(n^{-1/2}) \\ &= O(n^{-1/2}) \end{aligned}$$

Hence the order of the error is $O(n^{-1/2})$.

The good news is that the percentile-type methods (if appropriately used) can yield $\sqrt{n}$-convergent asymptotic inference. Yet these methods do not require the calculation of standard

errors! This means that in contexts where standard errors are not available or are difficult to calculate, the percentile bootstrap methods provide an attractive inference method.

The bad news is that the rate of convergence is disappointing. It is no better than the rate obtained from an asymptotic one-sided confidence region. Therefore if standard errors are available, it is unclear if there are any benefits from using the percentile bootstrap over simple asymptotic methods.

Based on these arguments, the theoretical literature (e.g. Hall, 1992, Horowitz, 2001) tends to advocate the use of the percentile-t bootstrap methods rather than percentile methods.

10.12 Bootstrap Methods for Regression Models

The bootstrap methods we have discussed have set $G_n^*(u) = G_n(u, F_n)$, where F_n is the EDF. Any other consistent estimate of F may be used to define a feasible bootstrap estimator. The advantage of the EDF is that it is fully nonparametric, it imposes no conditions, and works in nearly any context. But since it is fully nonparametric, it may be inefficient in contexts where more is known about F. We discuss bootstrap methods appropriate for the linear regression model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned}$$

The non-parametric bootstrap resamples the observations $(y_i^*, \boldsymbol{x}_i^*)$ from the EDF, which implies

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i^{*\prime}\hat{\boldsymbol{\beta}} + e_i^* \\ \mathbb{E}\left(\boldsymbol{x}_i^* e_i^*\right) &= \mathbf{0} \end{aligned}$$

but generally

$$\mathbb{E}\left(e_i^* \mid \boldsymbol{x}_i^*\right) \neq 0.$$

The bootstrap distribution does not impose the regression assumption, and is thus an inefficient estimator of the true distribution (when in fact the regression assumption is true.)

One approach to this problem is to impose the very strong assumption that the error ε_i is independent of the regressor $\boldsymbol{x}_i$. The advantage is that in this case it is straightforward to construct bootstrap distributions. The disadvantage is that the bootstrap distribution may be a poor approximation when the error is not independent of the regressors.

To impose independence, it is sufficient to sample the $\boldsymbol{x}_i^*$ and e_i^* independently, and then create $y_i^* = \boldsymbol{x}_i^{*\prime}\hat{\boldsymbol{\beta}} + e_i^*$. There are different ways to impose independence. A non-parametric method is to sample the bootstrap errors e_i^* randomly from the OLS residuals $\{\hat{e}_1, ..., \hat{e}_n\}$. A parametric method is to generate the bootstrap errors e_i^* from a parametric distribution, such as the normal $e_i^* \sim \mathrm{N}(0, \hat{\sigma}^2)$.

For the regressors $\boldsymbol{x}_i^*$, a nonparametric method is to sample the $\boldsymbol{x}_i^*$ randomly from the EDF or sample values $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. A parametric method is to sample $\boldsymbol{x}_i^*$ from an estimated parametric distribution. A third approach sets $\boldsymbol{x}_i^* = \boldsymbol{x}_i$. This is equivalent to treating the regressors as *fixed in repeated samples*. If this is done, then all inferential statements are made conditionally on the observed values of the regressors, which is a valid statistical approach. It does not really matter, however, whether or not the $\boldsymbol{x}_i$ are really "fixed" or random.

The methods discussed above are unattractive for most applications in econometrics because they impose the stringent assumption that $\boldsymbol{x}_i$ and e_i are independent. Typically what is desirable is to impose only the regression condition $\mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0$. Unfortunately this is a harder problem.

One proposal which imposes the regression condition without independence is the *Wild Bootstrap*. The idea is to construct a conditional distribution for e_i^* so that

$$\begin{aligned} \mathbb{E}\left(e_i^* \mid \boldsymbol{x}_i\right) &= 0 \\ \mathbb{E}\left(e_i^{*2} \mid \boldsymbol{x}_i\right) &= \hat{e}_i^2 \\ \mathbb{E}\left(e_i^{*3} \mid \boldsymbol{x}_i\right) &= \hat{e}_i^3. \end{aligned}$$

A conditional distribution with these features will preserve the main important features of the data. This can be achieved using a two-point distribution of the form

$$\begin{aligned} \Pr\left(e_i^* = \left(\frac{1+\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \Pr\left(e_i^* = \left(\frac{1-\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}+1}{2\sqrt{5}} \end{aligned}$$

For each $\boldsymbol{x}_i$, you sample e_i^* using this two-point distribution.

Exercises

Exercise 10.1 Let $F_n(\boldsymbol{x})$ denote the EDF of a random sample. Show that

$$\sqrt{n}\left(F_n(\boldsymbol{x}) - F_0(\boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}\left(0, F_0(\boldsymbol{x})\left(1 - F_0(\boldsymbol{x})\right)\right).$$

Exercise 10.2 Take a random sample $\{y_1, ..., y_n\}$ with $\mu = \mathbb{E}y_i$ and $\sigma^2 = \mathrm{var}\left(y_i\right)$. Let the statistic of interest be the sample mean $T_n = \overline{y}_n$. Find the population moments $\mathbb{E}T_n$ and $\mathrm{var}\left(T_n\right)$. Let $\{y_1^*, ..., y_n^*\}$ be a random sample from the empirical distribution function and let $T_n^* = \overline{y}_n^*$ be its sample mean. Find the bootstrap moments $\mathbb{E}T_n^*$ and $\mathrm{var}\left(T_n^*\right)$.

Exercise 10.3 Consider the following bootstrap procedure for a regression of y_i on $\boldsymbol{x}_i$. Let $\hat{\boldsymbol{\beta}}$ denote the OLS estimator from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, and $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the OLS residuals.

(a) Draw a random vector $(\boldsymbol{x}^*, e^*)$ from the pair $\{(\boldsymbol{x}_i, \hat{e}_i) : i = 1, ..., n\}$. That is, draw a random integer i' from $[1, 2, ..., n]$, and set $\boldsymbol{x}^* = \boldsymbol{x}_{i'}$ and $e^* = \hat{e}_{i'}$. Set $y^* = \boldsymbol{x}^{*\prime}\hat{\boldsymbol{\beta}} + e^*$. Draw (with replacement) n such vectors, creating a random bootstrap data set $(\boldsymbol{y}^*, \boldsymbol{X}^*)$.

(b) Regress $\boldsymbol{y}^*$ on $\boldsymbol{X}^*$, yielding OLS estimates $\hat{\boldsymbol{\beta}}^*$ and any other statistic of interest.

Show that this bootstrap procedure is (numerically) *identical* to the non-parametric bootstrap.

Exercise 10.4 Consider the following bootstrap procedure. Using the non-parametric bootstrap, generate bootstrap samples, calculate the estimate $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = (\hat{\theta}^* - \hat{\theta})/s(\hat{\theta}),$$

where $s(\hat{\theta})$ is the standard error in the original data. Let $q_n^*(.05)$ and $q_n^*(.95)$ denote the 5% and 95% quantiles of T_n^*, and define the bootstrap confidence interval

$$C = \left[\hat{\theta} - s(\hat{\theta})q_n^*(.95), \quad \hat{\theta} - s(\hat{\theta})q_n^*(.05)\right].$$

Show that C exactly equals the Alternative percentile interval (not the percentile-t interval).

Exercise 10.5 You want to test $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta > 0$. The test for $\mathbb{H}_0$ is to reject if $T_n = \hat{\theta}/s(\hat{\theta}) > c$ where c is picked so that Type I error is α. You do this as follows. Using the non-parametric bootstrap, you generate bootstrap samples, calculate the estimates $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = \hat{\theta}^*/s(\hat{\theta}^*).$$

Let $q_n^*(.95)$ denote the 95% quantile of T_n^*. You replace c with $q_n^*(.95)$, and thus reject $\mathbb{H}_0$ if $T_n = \hat{\theta}/s(\hat{\theta}) > q_n^*(.95)$. What is wrong with this procedure?

Exercise 10.6 Suppose that in an application, $\hat{\theta} = 1.2$ and $s(\hat{\theta}) = .2$. Using the non-parametric bootstrap, 1000 samples are generated from the bootstrap distribution, and $\hat{\theta}^*$ is calculated on each sample. The $\hat{\theta}^*$ are sorted, and the 2.5% and 97.5% quantiles of the $\hat{\theta}^*$ are .75 and 1.3, respectively.

(a) Report the 95% Efron Percentile interval for θ.

(b) Report the 95% Alternative Percentile interval for θ.

(c) With the given information, can you report the 95% Percentile-t interval for θ?

Exercise 10.7 The datafile `hprice1.dat` contains data on house prices (sales), with variables listed in the file `hprice1.pdf`. Estimate a linear regression of price on the number of bedrooms, lot size, size of house, and the colonial dummy. Calculate 95% confidence intervals for the regression coefficients using both the asymptotic normal approximation and the percentile-t bootstrap.