These notes come from Chapter 10 of the textbook and provide an introduction to resampling methods for conducting inference, particularly the bootstrap.

Resampling Methods

H: 10.1

Our approach to inference so far has been to establish the limiting distribution of some parameter of interest; for example, $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V}_\theta)$, and then to construct an estimate of $\mathbf{V}_\theta$. Given this estimate, we could construct a test statistic, for example a t-statistic for some $\mathbb{H}_0$, or construct a confidence interval, etc.

The idea of the resampling methods that we'll study in this section are, essentially, to substitute computational power for the (potentially complex) mathematical calculations that we have been using before. Resampling methods are popular in many applications. For example, the bootstrap is popular in quantile regression applications (which we'll talk about if we have time this semester) where (i) it is relatively complicated to figure out the asymptotic distribution and (ii) even after you derive the asymptotic distribution, it is relatively hard to estimate it.

The book talks briefly about two resampling methods that I'll just briefly mention here. The **jackknife** is the distribution from n leave-one-out estimators (e.g., taking turns estimating θ using all observations except one). **Sub-sampling** is like the boostrap that we'll talk about below except that you draw subsamples of the original data (with less than n observations) without replacement.

The Bootstrap Algorithm

H: 10.6

There are several variations of the bootstrap, but let's start with the most common one, which is typically called either the **nonparametric bootstrap** or the **empirical bootstrap**.

Step 1: Construct a **bootstrap sample** by making n iid draws, with replacement, from the original sample. We'll denote particular draws by (Y_i^*, X_i^*), and the entire bootstrap sample by $\{Y_i^*, X_i^*\}_{i=1}^n$.

Step 2: Construct the bootstrap estimate $\widehat{\theta}^*$ by applying whatever approach you originally used to estimate $\widehat{\theta}$ to the bootstrap sample. For example, if you are interested in the linear projection model, you would estimate $\widehat{\beta}^*$ by the linear regression of Y_i^* on X_i^*.

Steps 1 and 2 give us an estimate from the distribution of estimates obtained by iid sampling from the original data. However, the real usefulness of the bootstrap, is that (unlike our original sample from the population), we can repeat this process a large number of times. In particular, let B denote the number of bootstrap samples that we draw; then, for $b = 1, \ldots, B$, we can draw new bootstrap samples and calculate $\widehat{\theta}_b^*$, where the subscript indicates that it is the bootstrap estimate from the b^{th} bootstrap sample.

Other Types of Bootstrap Procedures

The nonparametric bootstrap procedure above is the most common one, but there are other variations that are worth mentioning.

The **weighted bootstrap** involves perturbing (i.e., causing it to vary) the objective function for some particular estimation procedure. For example, if you were trying to estimate $\mathbb{E}[Y]$, the bootstrap estimate would be given by

$$\hat{\mu}^* = \arg\min_m \frac{1}{n}\sum_{i=1}^n w_i(Y_i - m)^2$$

where w_i are iid weights (in particular, they are weights that are independent of each other and independent of the original data) that satisfy $\mathbb{E}[w] = 1$ and $\text{var}(w) = 1$. A leading choice is to make iid draws from an exponential distribution with mean 1 (in R, you can run `rexp(n)`). After solving this, you would get

$$\hat{\mu}^* = \left(\frac{1}{n}\sum_{i=1}^n w_i\right)^{-1} \frac{1}{n}\sum_{i=1}^n w_i Y_i$$

Similarly, if you were to compute bootstrap estimates of β from a regression, it would amount to computing

$$\hat{\beta}^* = \arg\min_b \frac{1}{n}\sum_{i=1}^n w_i(Y_i - X_i'b)^2$$

If you solve this, you will get

$$\hat{\beta}^* = \left(\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n w_i X_i Y_i$$

Side-Comment: The nonparametric bootstrap is actually quite related to the weighted bootstrap. In fact, you can write, for example, a nonparametric bootstrap estimate of $\widehat{\beta}^*$ by

$$\widehat{\beta}^* = \left(\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n w_i X_i Y_i$$

which is the same expression as for the weighted bootstrap. In this case $(w_1, w_2, \ldots, w_n)$ are drawn from a multinomial distribution with parameter n and probabilities $(1/n, 1/n, \ldots, 1/n)$. These weights have mean 1, but they are not independent (for example, if the weight on the first observation is large, it implies that the weight on other units is more likely to be small).

Another common approach is the **multiplier bootstrap** (sometimes this is called the **score**

bootstrap). In this case, bootstrap draws are constructed by perturbing the "score"/"influence function" (i.e., the part of the asymptotically linear representation of the estimator). For example, if we go back to the regression setup, we would compute bootstrap estimates by

$$\hat{\beta}^* = \hat{\beta} + \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} w_i X_i \hat{e}_i$$

where w_i are iid weights with $\mathbb{E}[w] = 0$ (note that this is different from the weighted bootstrap) and $\text{var}(w) = 1$. Common choices are (i) $W \sim N(0,1)$ or (ii) $W = 1$ with probability $1/2$ and $W = -1$ with probability $1/2$.

There are other variations of the bootstrap that we'll not cover; if you are interested, H: 10.29 covers the wild bootstrap, which is another popular version of the bootstrap and is commonly used in the context of nonparametric regression.

Bootstrap Variance and Standard Errors

H: 10.7

Once we have a large number of bootstrap estimates, we can estimate features of the bootstrap distribution of $\hat{\theta}_b^*$. The **bootstrap estimate of the asymptotic variance** of $\hat{\theta}$ is given by

$$\hat{\mathbf{V}}_\theta^{boot} = \frac{1}{B} \sum_{b=1}^{B} n \left(\hat{\theta}_b^* - \bar{\theta}^* \right) \left(\hat{\theta}_b^* - \bar{\theta}^* \right)'$$

where

$$\bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

When $\hat{\theta}$ is a scalar, the **bootstrap standard error** is given by

$$\widehat{\text{s.e.}}_{\hat{\theta}}^{boot} = \frac{\sqrt{\hat{\mathbf{V}}_{\hat{\theta}}^{boot}}}{\sqrt{n}}$$

As in the previous set of notes, it would be very common in applications to report $\hat{\theta}$ and $\widehat{\text{s.e.}}_{\hat{\theta}}^{boot}$. Moreover, bootstrap standard errors can be used to construct confidence intervals; e.g.,

$$C^{nb} = \left[\hat{\theta} \pm 1.96 \ \widehat{\text{s.e.}}_{\hat{\theta}}^{boot} \right]$$

where (I think) "nb" stands for "normal approximation bootstrap" (and comes from the notation in the textbook).

As an additional comment, although one would typically choose B to be a large number, it is still finite. This means that all bootstrap statistics, e.g., $\hat{\mathbf{V}}_\theta^{boot}$ are estimates and therefore are random.

In particular, this means that its value will change if you were to compute them more than once. This is to be expected, though typically they should be "close" if you were to compute them more than once.

The Bootstrap Distribution

H: 10.9

The remaining question that we should answer is: Why does the bootstrap work?

The book mainly talks about the nonparametric bootstrap. I strongly recommend reading H 10.9 which provides an explanation for the reason why the nonparametric bootstrap works. To very briefly summarize these arguments: First, our inference procedures come down to learning about the sampling distribution of our estimator, e.g., $\hat{\theta}$. The validity of the bootstrap basically comes down to, $\hat{F}$ (the empirical cdf of the observed data) should be "close" to F (the actual cdf of (Y, X)), and this approximation should get better for large n. The idea of the bootstrap is to (i) sample from $\hat{F}$ and then (ii) repeatedly simulate from this distribution. Given a large sample, this should be "similar" to repeatedly sampling from the population.

Bootstrap Asymptotics

H: 10.12

I am going to focus on understanding why the bootstrap works for the weighted bootstrap, as I think this is slightly easier to understand than the nonparametric bootstrap.

What we would like to show is something like that, at least asymptotically, $\sqrt{n}(\hat{\mu}^* - \hat{\mu})$ follows the same limiting distribution as $\sqrt{n}(\hat{\mu} - \mu)$ or that $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ follows the same limiting distribution as $\sqrt{n}(\hat{\beta} - \beta)$. This would provide a justification for, say, repeatedly calculating $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ (which is feasible) and using its distribution to learn about the distribution of $\sqrt{n}(\hat{\beta} - \beta)$.

To start with, an important conceptual issue worth emphasizing is that, when we construct bootstrap estimates, the data is fixed; what is random are the weights w_i. Thus, we will be interested in features of the distribution of bootstrap estimates *conditional on the observed data*. I'll follow the notation used in the textbook and denote expecations/variances conditional on the observed data by $\mathbb{E}^*$ and var^*. And, just to give an example, when we condition on the data, $\mathbb{E}^*[w_i Y_i] = \mathbb{E}[w] Y_i = Y_i$. In other words, (i) conditioning on the observed data, we treat Y_i as a constant so that it can come out of the expectation, (ii) w_i is independent of the observed data and identically distributed with mean one which implies that $\mathbb{E}^*[w_i] = \mathbb{E}[w_i] = 1$. Similarly, $\text{var}^*(w_i Y_i) = \text{var}(w_i) Y_i^2 = Y_i^2$.

There are bootstrap versions of all our key asymptotic tools: the law of large numbers, the central limit theorem, the continuous mapping theorem, and the delta method. This often means that establishing the validity of a bootstrap procedure is similar to establishing the limiting distribution of the estimator. I am just going to heuristically explain the bootstrap version of the weak law of large numbers and central limit theorem next.

Bootstrap WLLN Let $\bar{Y}^* = \frac{1}{n}\sum_{i=1}^{n} w_i Y_i$, and consider *iid* weights w satisfying $\mathbb{E}[w] = 1$ and $\text{var}(w) = 1$. If Y_i are iid and $\mathbb{E}|Y| < \infty$, then $\bar{Y}^* \xrightarrow{p^*} \mathbb{E}[Y]$ where $\xrightarrow{p^*}$ denotes "convergence in bootstrap probability".

What is happening here is two things: first, given a large number of observations $\frac{1}{n}\sum_{i=1}^{n} w_i Y_i$ should be close to its mean (conditional on the data) of $\bar{Y}$; second $\bar{Y}$ should be close to $\mathbb{E}[Y]$. Taken together, these suggest that, given a large enough sample, $\bar{Y}^*$ should be close to $\mathbb{E}[Y]$.

The proof of the bootstrap WLLN is very similar to the proof of the WLLN (see Theorem 10.2 in the textbook). The main difference is that, conditional on the observed data, $w_i Y_i$ are independent but not identically distributed (as discussed above, $\mathbb{E}^*[w_i Y_i] = Y_i$ and $\text{var}^*(w_i Y_i) = Y_i^2$, which implies that the conditional distribution of $w_i Y_i$ depends on Y_i). That said, this is not too challenging to deal with and arguments that mainly use Chebyshev's inequality go through here.

Bootstrap CLT Consider iid weights w that satisfy $\mathbb{E}[w] = 1$ and $var(w) = 1$, if Y_i are iid, $\mathbb{E}||Y||^2 < \infty$, and $\mathbf{\Sigma} := \text{var}(Y) > 0$, then $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} w_i(Y_i - \bar{Y})\right) \xrightarrow{d^*} N(0, \mathbf{\Sigma})$ where $\xrightarrow{d^*}$ denotes "convergence in bootstrap distribution".

Notice that the bootstrap CLT centers at $\bar{Y}$ rather than $\mathbb{E}[Y]$. I am not going to go into much technical detail here, but let me sketch the argument for this result. At first glance, the result for the bootstrap CLT may seem somewhat surprising especially because $\mathbb{E}^*[w_i(Y_i - \bar{Y})] = (Y_i - \bar{Y})$ which, in general, is not equal to 0 (and seems to suggest that we ought to be careful trying to apply some type of CLT here). However, notice that $0 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})$. This implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(Y_i - \bar{Y}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(Y_i - \bar{Y}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \bar{Y}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(w_i - 1)(Y_i - \bar{Y})$$

and now notice that $\mathbb{E}^*[(w_i - 1)(Y_i - \bar{Y})] = \mathbb{E}[(w-1)](Y_i - \bar{Y}) = 0$. This suggests that we might be able to apply a central limit theorem to this sort of term. As for the bootstrap WLLN, the main complication is that $(w_i - 1)(Y_i - \bar{Y})$ is not identically distributed; $\text{var}^*((w_i - 1)(Y_i - \bar{Y})) = \text{var}(w-1)(Y_i - \bar{Y})^2 = (Y_i - \bar{Y})^2$, which is not constant across i. This means that we cannot use the "Lindeberg-Levy" CLT that we have often used. Instead, however, we can use what the textbook calls the "Lindeberg" CLT. This CLT allows for independent but not identically distributed observations though it comes at the cost of requiring extra technical conditions (see Section 9.2 in Bruce Hansen's probability and statistics textbook for more details). Here, the Lindeberg CLT implies that $\hat{\mathbf{\Sigma}}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(Y_i - \bar{Y}) \xrightarrow{d^*} N(0, \mathbf{I})$ [where the multiplication by $\hat{\mathbf{\Sigma}}^{-1/2}$ arises because $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \text{var}^*((w_i - 1)(Y_i - \bar{Y})) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$]. Then, because $\hat{\mathbf{\Sigma}} \xrightarrow{p} \mathbf{\Sigma}$, it implies

the result.

The above explanation is mathematical, so let me take one more paragraph and explain the intuition for the bootstrap CLT in words. Like the Bootstrap WLLN, the right intuition to have here is a sort of two part argument. First, as $n \to \infty$, $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} w_i(Y_i - \bar{Y})\right)$ should behave like a draw from $N(0, \widehat{\text{var}}(Y))$; second $\widehat{\text{var}}(Y)$ should get close to $\text{var}(Y)$ as $n \to \infty$. Thus, in large samples, $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} w_i(Y_i - \bar{Y})\right)$ should behave like a draw from a $N(0, \mathbf{\Sigma})$ distribution. This is potentially useful because (i) it is the same distribution as $\sqrt{n}(\bar{Y} - \mathbb{E}[Y])$ follows, and (ii) we can use simulation to repeatedly calculate $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} w_i(Y_i - \bar{Y})\right)$.

Let's conclude this section by explaining why the weighted bootstrap works for approximating the limiting distribution of $\sqrt{n}(\hat{\mu} - \mu)$ (as we mentioned hoping for earlier in the notes). To start with, notice that

$$\begin{aligned}\hat{\mu}^* &= \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} w_i Y_i \\ &= \hat{\mu} + \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} w_i (Y_i - \bar{Y})\end{aligned}$$

where the second line uses $Y_i = \hat{\mu} + (Y_i - \bar{Y})$ (sorry for mixing notation: notice that $\hat{\mu} = \bar{Y}$ so this equation is a trivial one). This implies that

$$\sqrt{n}(\hat{\mu}^* - \hat{\mu}) = \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i (Y_i - \bar{Y}) \xrightarrow{d^*} N(0, \mathbf{\Sigma})$$

where the convergence result holds because (i) $\frac{1}{n}\sum_{i=1}^{n} w_i \xrightarrow{p^*} \mathbb{E}[W] = 1$, (ii) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i (Y_i - \bar{Y}) \xrightarrow{d^*} N(0, \mathbf{\Sigma})$ by the bootstrap CLT, (iii) combining these terms using the bootstrap CMT (which we didn't actually discuss above but works the same way as the CMT that we are used to).

Bootstrap Regression Asymptotic Theory

H: 10.28

To conclude this section, let's consider why the bootstrap works for approximating the limiting

distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ where $\hat{\beta}$ comes from the regression of Y on X. Recall that

$$\begin{aligned}
\hat{\beta}^* &= \left(\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n w_i X_i Y_i \\
&= \left(\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n w_i X_i (X_i\hat{\beta} + \hat{e}_i) \\
&= \hat{\beta} + \left(\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n w_i X_i \hat{e}_i
\end{aligned}$$

which implies that

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) = \left(\frac{1}{n}\sum_{i=1}^n w_i X_i X_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n w_i X_i \hat{e}_i$$

First, from the bootstrap WLLN, it follows that

$$\frac{1}{n}\sum_{i=1}^n w_i X_i X_i' \xrightarrow{p^*} \mathbb{E}[XX']$$

where the intuition is that (i) $n^{-1}\sum_{i=1}^n w_i X_i X_i'$ converges to its "population" mean $n^{-1}\sum_{i=1}^n X_i X_i'$ and (ii) $n^{-1}\sum_{i=1}^n X_i X_i$ converges to the actual population mean $\mathbb{E}[XX']$.

Second, notice that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i X_i \hat{e}_i = \frac{1}{\sqrt{n}}\sum_{i=1}^n w_i \left(X_i \hat{e}_i - n^{-1}\sum_{i=1}^n X_i \hat{e}_i\right) \xrightarrow{d^*} \mathbf{\Omega} = \mathbb{E}[XX'e^2]$$

where the first equality holds because $\frac{1}{n}\sum_{i=1}^n X_i \hat{e}_i = 0$ which is a property of running a regression (and I included this just to make it clear that we can apply the bootstrap CLT to this term), and the convergence in bootstrap distribution holds by the bootstrap CLT. Further, from the bootstrap continuous mapping theorem, we have that

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \xrightarrow{d^*} N(0, \mathbf{V})$$

where $\mathbf{V} = \mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1}$ which is the same as the limiting distribution for $\sqrt{n}(\hat{\beta} - \beta)$.