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, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V}_\theta)$, 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 $\hat{\theta}^*$ by applying whatever approach you originally used to estimate $\hat{\theta}$ to the bootstrap sample. For example, if you are interested in the linear projection model, you would estimate $\hat{\beta}^*$ 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 $\hat{\theta}_b^*$, 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 $\mathrm{E}[Y]$, 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 $\mathrm{E}[w] = 1$ and $\mathrm{var}(w) = 1$. 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 $\hat{\mu}^* = \frac{1}{n}\sum_{i=1}^n w_i Y_i$.

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 $\hat{\beta}^*$ by

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

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 $\mathrm{E}[w] = 0$ (note that this is different from the weighted bootstrap) and $\mathrm{var}(w) = 1$. 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

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

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

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

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.

The Distribution of Bootstrap Observations

H: 10.10

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.

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*.

Let's start by calculating the mean and variance of Y^* conditional on the data in the sample. Following the textbook, we will write these as, for example, $\mathrm{E}^*[Y^*]$, but this is the expected value of a draw of Y where we treat the sample as the population multiplied by w. Notice that

$$\mathrm{E}^*[Y^*] = \mathrm{E}^*[wY] = \mathrm{E}[w]\mathrm{E}^*[Y] = 1 \times \bar{Y} = \bar{Y}$$

where the second equality holds because w is independent of the original data, and the third equality holds because the mean of Y is $\bar{Y}$ when we treat the sample as the population. Thus, the (conditional on the original data) mean of Y^* is $\bar{Y}$. In particular, notice that it is not equal to $\mathrm{E}[Y]$. That said, this should not be surprising, because we are treated the data as the population. Similarly,

$$\mathrm{var}^*(Y^*) = \mathrm{E}^*[{Y^*}^2] - \mathrm{E}^*[Y^*]^2 = \mathrm{E}^*[wY^2] - \bar{Y}^2 = \mathrm{E}[w]\mathrm{E}^*[Y^2] - \bar{Y}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2 = \widehat{\mathrm{var}}(Y)$$

where the first equality holds by the definition of variance, the second equality holds because a weighted bootstrap draw of ${Y^*}^2$ amounts to making a draw of Y from the sample, squaring it, and multiplying it by a draw of w (this was the part where there was some confusion in class), the third equality uses independence of w, and the last equality holds because $\mathrm{E}^*[Y^2]$ is equal to the sample second moment of Y (due to treating the sample as the population). Thus, the variance of a weighted bootstrap draw of Y^* is given by $\widehat{\mathrm{var}}(Y)$ — the *estimated* variance of Y. Again, this is not surprising because we are treating the sample like it is the population.

The Distribution of the Bootstrap Sample Mean

H: 10.11

Using similar arguments as above (try these for practice), for $\bar{Y}^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^* = \frac{1}{n}\sum_{i=1}^{n} w_i Y_i$, you can show that

$$\mathrm{E}^*[\bar{Y}^*] = \bar{Y}$$

and that

$$\mathrm{var}^*(\bar{Y}^*) = \frac{\widehat{\mathrm{var}}(Y)}{n}$$

These arguments are similar to ones that show that $\mathrm{E}[\bar{Y}] = \mathrm{E}[Y]$ and that $\mathrm{var}(\bar{Y}) = \mathrm{var}(Y)/n$.

Bootstrap Asymptotics

H: 10.12

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 If Y_i are iid and $\mathrm{E}|Y| < \infty$, then $\bar{Y}^* \xrightarrow{p^*} \mathrm{E}[Y]$ 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 $\mathrm{E}[Y]$. Taken together, these suggest that, given a large enough sample, $\bar{Y}^*$ should be close to $\mathrm{E}[Y]$.

Bootstrap CLT If Y_i are iid, $\mathrm{E}||Y||^2 < \infty$, and $\mathbf{\Sigma} := \mathrm{var}(Y) > 0$, then $\sqrt{n}(\bar{Y}^* - \bar{Y}) \xrightarrow{d^*} N(0, \mathbf{\Sigma})$ where $\xrightarrow{d^*}$ denotes "convergence in bootstrap distribution".

Notice that the bootstrap CLT centers at $\bar{Y}$ rather than $\mathrm{E}[Y]$. Like the Bootstrap WLLN, the right intuition to have here is that, first, as $n \to \infty$, $\sqrt{n}(\bar{Y}^* - \bar{Y})$ should behave like a draw from $N(0, \widehat{\mathrm{var}}(Y))$ (because we are treating the sample like the population); second $\widehat{\mathrm{var}}(Y)$ should get close to $\mathrm{var}(Y)$ as $n \to \infty$. Thus, in large samples, $\sqrt{n}(\bar{Y}^* - \bar{Y})$ should behave like a draw from a $N(0, \mathbf{\Sigma})$ distribution. This is potentially useful because (i) it is the same distribution as $\sqrt{n}(\bar{Y} - \mathrm{E}[Y])$ follows, and (ii) we can use simulation to make repeated draws from $\sqrt{n}(\bar{Y}^* - \bar{Y})$.

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

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

Given that $w_i X_i X_i'$ is iid and has mean (conditional on the original data) $\mathrm{E}^*[wXX'] = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ (which follows using similar arguments to the ones for $\bar{Y}^*$ above), it follows from bootstrap WLLN that

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

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 $\mathrm{E}[XX']$.

Similarly, notice that (conditional on the original data) $w_i X_i \hat{e}_i$ is iid with mean $\mathrm{E}^*[wX\hat{e}] = \frac{1}{n}\sum_{i=1}^n X_i \hat{e}_i = 0$, and with variance equal to $\mathrm{var}^*(wX\hat{e}) = \frac{1}{n}\sum_{i=1}^n X_i X_i' \hat{e}_i^2 = \hat{\mathbf{\Omega}}$. Thus, by the bootstrap CLT

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

and further from the bootstrap continuous mapping theorem (which we didn't actually discuss above but works the same way as the CMT that we are used to) that

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

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