These notes cover chapter 4 of the textbook. We will be considering the finite sample properties of the linear CEF model.

Linear Regression Notes 3: Least squares regression

H: 4.3, 4.4

We will consider the following assumptions throughout this part of the course:

1. Linear CEF: $Y = X'\beta + e$ and $\mathrm{E}[e|X] = 0$
2. Finite Moments: $\mathrm{E}[Y^2] < \infty$ and $\mathrm{E}||X||^2 < \infty$
3. Positive definite design matrix: $\mathrm{E}[XX']$ is positive definite.

For some of the results below, we will also use the additional **homoskedasticity** condition: $\mathrm{E}[e^2|X] = \sigma^2$ (that is, the variance of the error term does not depend on X)

We'll continue to suppose that we have access to an i.i.d. sample. The main two properties that we'll consider are the **bias** of $\hat{\beta}$ and the **sampling variance** of $\hat{\beta}$. Before we consider those, let's start by defining what they are. Let $\hat{\theta}$ generically denote some estimator of a population parameter of interest θ. Then,

$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta$$

$\hat{\theta}$ is said to be **unbiased** if $\mathrm{Bias}(\hat{\theta}) = 0$, or, equivalently, if $\mathrm{E}[\hat{\theta}] = \theta$. It is worth pausing a moment to think conceptually about what is happening here. First, estimators are random — this point may not be immediately obvious though. In particular, given once you have access to a particular dataset, this typically pins down a value of $\hat{\theta}$. What it means that $\hat{\theta}$ is random is that we can carry out the thought experiment of repeatedly collecting n new observations from the same population and re-calculating $\hat{\theta}$ for the new data. In our thought experiment, given that we have new samples, the value of $\hat{\theta}$ would generally change with each new sample. If you were to carry this procedure out an extremely large number of times, this would give rise to a distribution of $\hat{\theta}$ in repeated samples; this distribution is called the **sampling distribution** of $\hat{\theta}$.

In practice, however, we only have one dataset and, therefore, only one value of $\hat{\theta}$. Given the above discussion, it is natural to consider the $\hat{\theta}$ that we have as a draw from the sampling distribution discussed above. Therefore, if an estimator is unbiased, what this means is that, on average (with respect to the sampling distribution), our estimate $\hat{\theta}$ is equal to the population parameter θ. Importantly, unbiasedness is generally a good property for an estimator to have, but, given that we only have one draw from the sampling distribution, even if our estimator is unbiased, it is still *possible* that our particular value of $\hat{\theta}$ could be far away from θ.

Practice: Show that $\bar{Y} := \frac{1}{n}\sum_{i=1}^n Y_i$ is unbiased for $\mathrm{E}[Y]$.

Next, the sampling variance of $\hat{\theta}$ is given by $\text{var}(\hat{\theta})$. You should think of this as the variance of $\hat{\theta}$ in the repeated sampling thought experiment mentioned above. All else equal, we would prefer estimators that have lower sampling variance.

Expectation of least squares estimator

H: 4.5, 4.7

Now, let's consider the bias of $\hat{\beta}$. To start with let's calculate $\mathrm{E}[\hat{\beta}|\mathbf{X}]$ (this sort of conditional expectation may feel a bit unusual as we are conditioning on the data matrix, but it is totally reasonable to do this)

$$\begin{aligned}
\mathrm{E}[\hat{\beta}|\mathbf{X}] &= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\mathbf{Y}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\
&= \beta
\end{aligned}$$

To see the step that uses $\mathrm{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$, let's point out a few things. First,

$$\mathrm{E}[Y_i|\mathbf{X}] = \mathrm{E}[Y_i|X_1, X_2, \ldots, X_n] = \mathrm{E}[Y_i|X_i] = X_i'\beta$$

where the first equality holds immediately, the second equality holds by the independence in i.i.d. sampling, and the last equality holds by the linear CEF. Thus,

$$\mathrm{E}[\mathbf{Y}|\mathbf{X}] = \begin{pmatrix} \vdots \\ \mathrm{E}[Y_i|\mathbf{X}] \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ X_i'\beta \\ \vdots \end{pmatrix} = \mathbf{X}\beta$$

which is what we used above.

The book provides an alternative derivation for the same result which I think is also useful for quickly covering. Notice that we can alternatively write

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{X}\beta + \mathbf{e})) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{e} \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{e} \qquad (1)
\end{aligned}$$

The expression in Equation 1 is one that we will use a number of times throughout this semester, so I think it is worth highlighting.

Now, using this expression, notice that

$$\begin{aligned}\mathrm{E}[\hat{\beta}|\mathbf{X}] &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathrm{E}[\mathbf{e}|\mathbf{X}] \\ &= \beta\end{aligned}$$

where the last equality holds because $\mathrm{E}[\mathbf{e}|\mathbf{X}] = \mathbf{0}$ which holds because $\mathrm{E}[e|X] = 0$ and by using similar arguments as for $\mathrm{E}[\mathbf{Y}|\mathbf{X}]$ above.

Given the result above, it then follows by the law of iterated expectations that

$$\mathrm{E}[\hat{\beta}] = \mathrm{E}\big[\mathrm{E}[\hat{\beta}|\mathbf{X}]\big] = \beta$$

and that, therefore, $\hat{\beta}$ is unbiased for β.

Variance of least squares estimator

H: 4.6, 4.7

Next, we'll calculate the sampling variance of $\hat{\beta}$. To this end, let's start by defining

$$\mathbf{D} := \mathrm{var}(\mathbf{e}|\mathbf{X}) = \mathrm{E}[\mathbf{e}\mathbf{e}'|\mathbf{X}]$$

where the last equality holds because $\mathrm{E}[\mathbf{e}|\mathbf{X}] = \mathbf{0}$. It's worth momentarily thinking about some of the properties of $\mathbf{D}$. First, it is an $n \times n$ matrix. Second, it's diagonal elements are given by $\mathrm{E}[e_i^2|\mathbf{X}] = \mathrm{E}[e_i^2|X_i] =: \sigma_i^2$. The off-diagonal elements are given by $\mathrm{E}[e_ie_j|\mathbf{X}] = \mathrm{E}[e_i|X_i]\mathrm{E}[e_j|X_j] = 0$ (here, the second equality holds by independence across observations). Thus, $\mathbf{D}$ is a diagonal matrix. If we are willing to introduce the assumption of homoskedasticity, then $\mathrm{E}[e_i^2|X_i] = \sigma^2$ (and is therefore constant across i). In this case, $\mathbf{D} = \boldsymbol{I}_n\sigma^2$.

Now, consider

$$\begin{aligned}\mathbf{V}_{\hat{\beta}} &:= \mathrm{var}(\hat{\beta}|\mathbf{X}) \\ &= \mathrm{var}\big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}\big) \\ &= \mathrm{var}\big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e})|\mathbf{X}\big) \\ &= \mathrm{var}\big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|\mathbf{X}\big) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{var}(\mathbf{e}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

where the fourth equality holds because the conditional variance of the part involving $\mathbf{X}\beta$ is equal to 0, the fifth equality by the vector version of $\mathrm{var}(aZ) = a^2\mathrm{var}(Z)$ when a is a constant and Z is a scalar random variable (and because $\mathbf{X}'\mathbf{X}$ is symmetric), and the last equality holds by the definition of $\mathbf{D}$. If we additionally invoke homoskedasticity, then this will simplify; in particular, in this case $\mathbf{X}'\mathbf{D}\mathbf{X} = \mathbf{X}'\boldsymbol{I}_n\sigma^2\mathbf{X} = \mathbf{X}'\mathbf{X}\sigma^2$. This implies that

$$\mathbf{V}_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

If we want to calculate the unconditional variance of $\hat{\beta}$, then we can use the law of total variance. This is given in Theorem 2.8 in the textbook; in particular, as along as $\mathrm{E}[Y^2] < \infty$, then $\mathrm{var}(Y) = \mathrm{E}[\mathrm{var}(Y|X)] + \mathrm{var}(\mathrm{E}[Y|X])$. Applying this to the present context, we have that

$$\begin{aligned}
\mathrm{var}(\hat{\beta}) &= \mathrm{E}[\mathrm{var}(\hat{\beta}|\mathbf{X})] + \mathrm{var}(\mathrm{E}[\hat{\beta}|\mathbf{X}]) \\
&= \mathrm{E}[\mathrm{var}(\hat{\beta}|\mathbf{X})] + 0 \\
&= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]
\end{aligned}$$

as above, this can simplify under homoskedasticity.

Gauss-Markov Theorem

H: 4.8

The Gauss-Markov theorem says that, given the linear regression assumptions + homoskedasticity, $\hat{\beta}$ is **efficient** (has the smallest variance) among all possible *linear*, *unbiased* estimators (side-comment: Bruce Hansen has a recent paper showing that $\hat{\beta}$ is efficient among unbiased estimators; I am not sure that I fully understand his arguments, so I'm just going to teach the "classical' ' version of the Gauss-Markov theorem).

More specifically, the Gauss-Markov theorem says: Given the linear regression assumptions and homoskedasticity, for any possible linear, unbiased estimator of β, which we'll denote as $\tilde{\beta}$, $\mathrm{var}(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Efficiency is a very good property for an estimator to have, and, therefore, this kind of result provides a strong justification for using $\hat{\beta}$ as an estimate of β.

To prove this result, let's first see what linearity and unbiasedness "buy us' '.

1. A linear estimator is one that we can write as $\tilde{\beta} = \mathbf{A}'\mathbf{Y}$ where $\mathbf{A}$ is an $n \times k$ matrix that is a function of $\mathbf{X}$

2. Unbiasedness means that $\mathrm{E}[\tilde{\beta}|\mathbf{X}] = \beta$. If $\tilde{\beta}$ is also linear, notice that $\mathrm{E}[\mathbf{A}'\mathbf{Y}|\mathbf{X}] = \mathbf{A}'\mathrm{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\beta$; then, unbiasedness therefore implies that $\mathbf{A}'\mathbf{X} = \boldsymbol{I}_k$.

Now, let's calculate the conditional variance of some generic linear, unbiased estimator of β

$$\begin{aligned}
\text{var}(\tilde{\beta}|\mathbf{X}) &= \text{var}(\mathbf{A}'\mathbf{Y}|\mathbf{X}) \\
&= \text{var}(\mathbf{A}'(\mathbf{X}\beta + \mathbf{e})|\mathbf{X}) \\
&= \text{var}(\mathbf{A}'\mathbf{e}|\mathbf{X}) \\
&= \mathbf{A}'\text{var}(\mathbf{e}|\mathbf{X})\mathbf{A} \\
&= \mathbf{A}'\mathbf{A}\sigma^2
\end{aligned}$$

where the first equality holds by linearity, the second equality substitutes for $\mathbf{Y}$, the third equality holds because the variance of the term involving $\mathbf{X}\beta$ is equal to 0 conditional on $\mathbf{X}$, the fourth equality holds by the property of variance that we used above (and because $\mathbf{A}$ is a function of $\mathbf{X}$), and the last equality holds because $\text{var}(\mathbf{e}|\mathbf{X}) = \boldsymbol{I}_n\sigma^2$ under homoskedasticity.

Since, from earlier, we know that $\text{var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, to complete the proof, we need to show that $\mathbf{A}'\mathbf{A} \geq (\mathbf{X}'\mathbf{X})^{-1}$. Towards this end, notice that

$$\begin{aligned}
\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} &= \mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A} \\
&= \mathbf{A}'(\boldsymbol{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{A} \\
&= \mathbf{A}'\mathbf{M}\mathbf{A} \\
&= \mathbf{A}'\mathbf{M}\mathbf{M}\mathbf{A} \\
&= \mathbf{A}'\mathbf{M}'\mathbf{M}\mathbf{A} \\
&= (\mathbf{M}\mathbf{A})'\mathbf{M}\mathbf{A} \\
&\geq 0
\end{aligned}$$

where the first equality uses $\mathbf{A}'\mathbf{X} = \boldsymbol{I}_k$, the second equality factors out $\mathbf{A}$, the third equality holds by the definition of $\mathbf{M}$, the fourth and fifth equalities hold because $\mathbf{M}$ is idempotent and symmetric, the term in the last equality is positive semi-definite because it is a quadratic form.

Generalized least squares

H: 4.9

The Gauss-Markov theorem relied on the homoskedasticity condition. This begs the question of whether or not these efficiency results for $\hat{\beta}$ go through without this condition. Section 4.9 of the book considers this case. In fact, it considers a more general case than we have been considering so far where $\text{var}(\mathbf{e}|\mathbf{X}) = \Sigma\sigma^2$ where Σ is an $n \times n$ symmetric and positive semi-definite matrix (what's more general here is that this allows for relaxing the independence condition so that Σ can be non-diagonal).

Using similar arguments as above, we can show that, in this case

$$\text{var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Sigma\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

However, Theorem 4.5 in the textbook shows that, under the linear regression assumptions (but not requiring homoskedasticity), for any possible linear, unbiased estimator of β (again, we'll denote it $\tilde{\beta}$),

$$\mathrm{var}(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$$

Since $\mathrm{var}(\hat{\beta}|\mathbf{X}) \neq \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$, this suggests that we might ought to consider alternative estimators in this case. In particular, when Σ is known, consider pre-multiplying the regression by $\Sigma^{-1/2}$ to get

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\beta + \tilde{\mathbf{e}}$$

where $\tilde{\mathbf{Y}} := \Sigma^{-1/2}\mathbf{Y}$, $\tilde{\mathbf{X}} := \Sigma^{-1/2}\mathbf{X}$, and $\tilde{\mathbf{e}} := \Sigma^{-1/2}\mathbf{e}$, and consider estimating this by OLS, so that

$$\begin{aligned}
\tilde{\beta}_{gls} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \\
&= \left((\Sigma^{-1/2}\mathbf{X})'\Sigma^{-1/2}\mathbf{X}\right)^{-1}(\Sigma^{-1/2}\mathbf{X})'\Sigma^{-1/2}\mathbf{Y} \\
&= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}
\end{aligned}$$

Using the same sorts of arguments as we have been making above, you can show the following two results

$$\begin{aligned}
\mathrm{E}[\tilde{\beta}_{gls}|\mathbf{X}] &= \beta \\
\mathrm{var}(\tilde{\beta}_{gls}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}
\end{aligned}$$

This suggests that $\tilde{\beta}_{gls}$ is both unbiased and more efficient that $\hat{\beta}$ under homoskedasticity.

One issue, however, is that this estimator is generally infeasible because Σ is not typically known. Instead, in practice, you can replace Σ with a suitable estimate $\hat{\Sigma}$. This is called **feasible GLS**. My sense is the GlS/FGLS is not very common in applied work, but I am not 100% the reason; the book does not talk very much about this. I checked Mostly Harmless Econometrics, and it recommends sticking with OLS because (i) estimating Σ may be hard to do in practice, replacing Σ with $\hat{\Sigma}$ might not work well in finite samples (i.e., realistic applications), (ii) the arguments in this section have been for the case where the CEF is actually linear, so it is less clear if there is a gain to using FGLS when we view $\hat{\beta}$ as the linear projection coefficient instead of the coefficient from a linear CEF model.