Supplementary Regression Notes

Bias and Variance of $\hat{\beta}$

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

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

For some of the results below, we will also use the additional **homoskedasticity** condition: $\mathbb{E}[e^2|X] = \sigma^2$ (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,

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

$\hat{\theta}$ is said to be **unbiased** if $\text{Bias}(\hat{\theta}) = 0$, or, equivalently, if $\mathbb{E}[\hat{\theta}] = \theta$. 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 estimator $\hat{\theta}$ 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 $\mathbb{E}[Y]$.

Next, the sampling variance of $\hat{\theta}$ is given by $\mathrm{var}(\hat{\theta})$. 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 $\mathbb{E}[\hat{\beta}|\mathbf{X}]$ (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)

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

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

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

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,

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

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

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

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

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

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

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

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

where the last equality holds because $\mathbb{E}[\mathbf{e}|\mathbf{X}] = \mathbf{0}$. 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 $\mathbb{E}[e_i^2|\mathbf{X}] = \mathbb{E}[e_i^2|X_i] =: \sigma_i^2$. The off-diagonal elements are given by $\mathbb{E}[e_i e_j|\mathbf{X}] = \mathbb{E}[e_i|X_i]\mathbb{E}[e_j|X_j] = 0$ (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 $\mathbb{E}[e_i^2|X_i] = \sigma^2$ (and is therefore constant across i). In this case, $\mathbf{D} = \mathbf{I}_n\sigma^2$.

As a first step towards calculating $\text{var}(\hat{\beta})$, notice that

$$\begin{aligned}\text{var}(\mathbf{Y}|\mathbf{X}) &= \text{var}(\mathbf{X}\beta + \mathbf{e}|\mathbf{X}) \\ &= \text{var}(\mathbf{e}|\mathbf{X}) = \mathbf{D}\end{aligned}$$

where the first equality holds by plugging in for $\mathbf{Y}$, the second equality holds because we are conditioning on $\mathbf{X}$, and the last equality by the definition of $\mathbf{D}$.

Now, consider

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

where the second equality holds by plugging in for $\hat{\beta}$, the third equality by the matrix 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 because $\mathrm{var}(\mathbf{Y}|\mathbf{X}) = \mathbf{D}$ which we showed above. If we additionally invoke homoskedasticity, then this will simplify; in particular, in this case $\mathbf{X}'\mathbf{D}\mathbf{X} = \mathbf{X}'\mathbf{I}_n\sigma^2\mathbf{X} = \mathbf{X}'\mathbf{X}\sigma^2$. This implies that

$$\mathbf{V}^0_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

where I include the 0 superscript to indicate that this expression holds only under the additional condition of homoskedasticity.

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 $\mathbb{E}[Y^2] < \infty$, then $\mathrm{var}(Y) = \mathbb{E}[\mathrm{var}(Y|X)] + \mathrm{var}(\mathbb{E}[Y|X])$. Applying this to the present context, we have that

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

as above, this can simplify under homoskedasticity.

Side-comment: It is worth briefly comparing the above results to similar results in the very simple case where we estimate $\mu := \mathbb{E}[Y]$ by $\bar{Y}$ (the sample average of Y_i). In this case, recall that $\mathbb{E}[\bar{Y}] = \mu$, so that $\bar{Y}$ is unbiased for μ, just like $\hat{\beta}$ is for β.

Further, recall that $\textrm{var}(\bar{Y}) = \frac{\textrm{var}(Y)}{n}$, which says that the sampling variance of $\bar{Y}$ depends on the variance of Y, and it also tends to decrease for larger values of n. From the above discussion, it may not be immediately obvious whether or not the sampling variance of $\hat{\beta}$ decreases with n — it turns out that it does. To see this, recall that $(\mathbf{X}'\mathbf{X}) = \sum_{i=1}^n X_i X_i'$ which grows with n. Now, for simplicity, suppose that homoskedasticity holds (similar arguments will hold for the case without homoskedasticity), notice that we can rewrite

$$\mathbf{V}^0_{\hat{\beta}} = \frac{\sigma^2}{n} \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1}$$

which just multiplies and divides by n. Notice that, here, $\displaystyle\frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ is now an average that does not systematically grow with n. On the other hand, there is now an n in the denominator so that it is easier to see that the sampling variance of $\hat{\beta}$ *does* decrease with the sample size, just like for $\bar{Y}$.

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.

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

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 "buys 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 $\mathbb{E}[\tilde{\beta}|\mathbf{X}] = \beta$. If $\tilde{\beta}$ is also linear, notice that $\mathbb{E}[\mathbf{A}'\mathbf{Y}|\mathbf{X}] = \mathbf{A}'\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\beta$; then, unbiasedness therefore implies that $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$.

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

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

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 $\mathrm{var}(\mathbf{e}|\mathbf{X}) = \mathbf{I}_n\sigma^2$ under homoskedasticity.

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

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

where the first equality uses $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$, 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 $\mathrm{var}(\mathbf{e}|\mathbf{X}) = \sigma^2\mathbf{\Sigma}$ where $\mathbf{\Sigma}$ is an $n \times n$ symmetric and positive definite matrix (what's more general here is that this allows for relaxing the independence condition so that $\mathbf{\Sigma}$ can be non-diagonal).

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

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

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 $\tilde{\beta}$),

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

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

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

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

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

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

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

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

One issue, however, is that this estimator is generally infeasible because $\mathbf{\Sigma}$ is not typically known. Instead, in practice, you can replace $\mathbf{\Sigma}$ with a suitable estimate $\hat{\mathbf{\Sigma}}$. This is called **feasible GLS**. My sense is that GLS/FGLS is not very common in applied work, especially relative to OLS combined with "heteroskedasticity robust" standard errors. I think there are several reasons for this. First, estimating $\mathbf{\Sigma}$ may be hard to do in practice. For example, if we return to the simpler case where $\mathrm{var}(\mathbf{e}|\mathbf{X}) = \mathbf{D}$ and recalling that $\mathbf{D}$ is diagonal with diagonal elements equal to $\mathbb{E}[e_i^2|X_i]$. To estimate $\mathbf{D}$ then would require estimating $\mathbb{E}[e^2|X]$. In practice, you could write down a parametric model for $\mathbb{E}[e^2|X]$, but this might be difficult in practice. If the model is not correctly specified, then the efficiency arguments above may not hold anymore. Second, the arguments that rationalize FGLS typically require $n \to \infty$ and amount to showing that FGLS and GLS are equivalent in this case (I think the finite sample arguments that we have been considering above for OLS/GLS are not straightforward when $\mathrm{var}(\mathbf{e}|\mathbf{X})$ has to be estimated). This somewhat weakens the positive results for GLS mentioned above. Finally, 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.

Frisch-Waugh-Lovell Theorem

H: 2.23

In this section, we will discuss the Frisch-Waugh-Lovell (FWL) Theorem. This is a very important result in econometrics as it (i) helps to provide a clearer interpretation of the parameters in a linear regression, especially in settings where we are particularly interested in some of the parameters and not others—this is quite common in economics where often we are interested in the effect of one variable on the outcome while trying to control for a large number of other variables. In particular, it provides a mechanical interpretation of what it means to control for other regressors. (ii) It is a useful first step in deriving several other results that we will talk about later in the semester. (iii) It is computationally useful in some cases such as some of the panel data approaches that we will consider later in the semester.

Suppose that we are going to estimate the following regression. This is the same regression that we have been considering except that we have partitioned X and β into two parts, where X_1 is a k_1 dimensional vector, X_2 is a k_2 dimensional vector, and $k = k_1 + k_2$. You can think of X_1 as being the regressors that we are particularly interested in and X_2 as being the regressors that we are less interested in but want to control for. Using this notation, we can write the linear projection of Y on X as

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

where $\mathbb{E}[Xe] = \begin{bmatrix} \mathbb{E}[X_1 e] \\ \mathbb{E}[X_2 e] \end{bmatrix} = 0$. Now, consider the auxiliary linear projections of Y on X_2 and X_1 on X_2, that is,

$$Y = X_2'\gamma_2 + u \quad \text{and} \quad X_1 = \Lambda_{12} X_2 + v$$

where $\mathbb{E}[X_2 u] = 0$ and $\mathbb{E}[vX_2'] = 0$ (this is a $k_1 \times k_2$ matrix, which is effectively saying that the projection error from each of the regressions of the k_1 elements of X_1 on X_2 is uncorrelated with X_2). Here, Λ_{12} is a $k_1 \times k_2$ matrix of coefficients from the linear projection of all k_1 elements of X_1 on X_2; notice that u is a scalar and v is a k_1 dimensional vector.

FWL Theorem:

In the setting given above, $\beta_1 = \mathbb{E}[vv']^{-1}\mathbb{E}[vu]$.

The FWL theorem says that β_1 is the coefficient from a regression of u on v, which is the same as a regression of Y on X_1 after "partialling out" X_2 (i.e., removing the linear relationship between Y and X_2 and the linear relationship between X_1 and X_2). This gives a mechanical interpretation of what it means to "control for" X_2 in a regression of Y on X_1 and X_2. Is this a "good" result for regressions being useful for controlling for other variables? In some ways yes and in some ways no, but this is a discussion that we will have in much more detail later on in the semester.

Proof: Proof of FWL Theorem Notice that

$$\begin{aligned}
\mathbb{E}[vu] &= \mathbb{E}[v(Y - X_2'\gamma_2)] \\
&= \mathbb{E}[vY] - \underbrace{\mathbb{E}[vX_2']}\gamma_2 \\
&= \mathbb{E}[v(X_1'\beta_1 + X_2'\beta_2 + e)] \\
&= \mathbb{E}[vX_1']\beta_1 + \underbrace{\mathbb{E}[vX_2']}\beta_2 + \mathbb{E}[ve] \\
&= \mathbb{E}[vX_1']\beta_1 + \mathbb{E}[ve] \\
&= \mathbb{E}[vX_1']\beta_1 + \underbrace{\mathbb{E}[(X_1 - \Lambda_{12}X_2)e]} \\
&= \mathbb{E}[vX_1']\beta_1
\end{aligned}$$

where all of the underlined terms are equal to 0 by properties of the projection errors, and the other lines follow by substituting for u, Y and v. Next, notice that

$$\begin{aligned}
\mathbb{E}[vv'] &= \mathbb{E}[v(X_1 - \Lambda_{12}X_2)'] \\
&= \mathbb{E}[vX_1'] - \underbrace{\mathbb{E}[vX_2']}\Lambda_{12}' \\
&= \mathbb{E}[vX_1']
\end{aligned}$$

which holds by the properties of projection errors and substituting. Putting the previous two results together, we have that

$$\beta_1 = \mathbb{E}[vv']^{-1}\mathbb{E}[vu]$$

Wald Statistic

A Wald statistic is probably the second most common test statistic in empirical work (after the t-statistic). It is used to test a multivariate null hypothesis. We will disuss Wald statistics in the context of linear regression, but they are more general and arise in other contexts where you are interested in testing a multivariate null hypothesis. In particular, suppose we are interested in testing a multivariate null hypothesis concerning the parameters of $Y = X'\beta + e$.

Examples

1. $\beta_1 = \beta_2 = \beta_3 = 0$

2. $\beta_1 = \beta_2 = \beta_3$

3. $\beta_1 + \beta_2 = 1$

Notice that you can write all of these null hypotheses in the form $H_0 : \mathbf{R}'\beta = r$ where $\mathbf{R}$ is a $k \times q$ matrix (q being the number of restrictions implied by the null hypothesis) and r is a q dimensional vector. In the examples above,

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{bmatrix} \quad r_1 = \mathbf{0}, \quad \mathbf{R}_2 = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \quad r_2 = \mathbf{0}, \quad \mathbf{R}_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad r_3 = 1$$

where, for example, $\mathbf{R}_2$ is the matrix that encodes the restrictions implied by the null hypothesis in example 2.

Most of our intuition from t-statistics will carry over to a Wald statistic: for a t-statistic, suppose that you were interested in $H_0 : \beta_1 = 0$. In this case, what we would do is to take our estimate $\hat{\beta}_1$ and try to determine if it is "close" to 0 or not—if it's not, then we reject the H_0; otherwise, we fail to reject the H_0. Here, we will do the same thing except that we will try to determine if $\mathbf{R}'\hat{\beta}$ is close to r or not.

Operationalizing this idea is slightly more complicated than for a t-statistic though because $\mathbf{R}'\hat{\beta}$ is a q dimensional vector. As a first step, let us consider the quantity $\sqrt{n}(\mathbf{R}'\hat{\beta} - r)$. If H_0 is true, then $\mathbf{R}'\beta - r = 0$, so that

$$\sqrt{n}(\mathbf{R}'\hat{\beta} - r) = \sqrt{n}\Big((\mathbf{R}'\hat{\beta} - r) - (\mathbf{R}'\beta - r)\Big) = \sqrt{n}(\mathbf{R}'\hat{\beta} - \mathbf{R}'\beta) = \mathbf{R}'\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Lambda})$$

If H_0 is not true, then $\mathbf{R}'\beta - r \neq 0$, which implies that $\mathbf{R}'\hat{\beta} \xrightarrow{p} \mathbf{R}'\beta \neq r$, further implying that $\sqrt{n}(\mathbf{R}'\hat{\beta} - r)$ diverges (or at least one element of it diverges).

Next, let us form a Wald statistic that can exploit the differing behavior of $\sqrt{n}(\mathbf{R}'\hat{\beta} - r)$ under the null and alternative hypotheses. In particular, we can form the following Wald statistic

$$W = n(\mathbf{R}'\hat{\beta} - r)'\hat{\mathbf{\Lambda}}^{-1}(\mathbf{R}'\hat{\beta} - r)$$

You can inuitively think of this as "squaring" $\sqrt{n}(\mathbf{R}'\hat{\beta} - r)$, where the squaring is weighted by $\hat{\mathbf{\Lambda}}^{-1}$, which is the estimated inverse variance of $\sqrt{n}(\mathbf{R}'\hat{\beta} - r)$, and then multiplying by n.

Next, let us formalize that W has different behavior depending on whether or not H_0 is true.

Case 1: H_0 is true

$$\begin{aligned}
W &= n(\mathbf{R}'\hat{\beta} - r)'\mathbf{\Lambda}^{-1}(\mathbf{R}'\hat{\beta} - r) + o_p(1) \\
&= \left(\mathbf{\Lambda}^{-1/2}\sqrt{n}(\mathbf{R}'\hat{\beta} - r)\right)'\left(\mathbf{\Lambda}^{-1/2}\sqrt{n}(\mathbf{R}'\hat{\beta} - r)\right) + o_p(1) \\
&\xrightarrow{d} Z'Z \sim \chi^2_q
\end{aligned}$$

where the first equality holds because $\hat{\mathbf{\Lambda}}$ is consistent for $\mathbf{\Lambda}$, the second equality holds by factoring out $\mathbf{\Lambda}^{-1/2}$, and the last line holds because $\mathbf{\Lambda}^{-1/2}\sqrt{n}(\mathbf{R}'\hat{\beta}-r) \xrightarrow{d} Z \sim \mathcal{N}(0, I_q)$ and because $Z'Z \sim \chi^2_q$.

Case 2: H_0 is not true

In this case, $\mathbf{R}'\hat{\beta} \xrightarrow{p} \mathbf{R}'\beta \neq r$, which implies that terms like $\sqrt{n}(\mathbf{R}'\hat{\beta} - r)$ diverge, which further implies that W diverges.

The discussion above implies different behavior under the null and alternative hypotheses. Under H_0, W should behave like a draw from a chi-squared distribution with q degrees of freedom, while under the alternative, W should diverge. This suggests that we can use W to test H_0 by comparing it to the appropriate critical values coming from a chi-squared distribution with q degrees of freedom, rejecting H_0 if W is larger than the critical value.

In practice, instead of reporting W itself or the result of comparing it to a critical value, it is common (and probably better practice as it is easier to understand) to report a p-value. In `R`, you can compute a p-value by

```
1 - pchisq(W, df = q)
```

where `W` is the value of the Wald statistic you computed.

Functions of Parameters

H: 7.10

In many applications, a researcher may only be interested in conducting inference with respect to a specific transformation of the parameters. Probably the leading case is when a researcher is just interested in a particular parameter, say, β_1; but another example would be a case where a researcher is interested in, say, β_j/β_l (the ratio between β_j and β_l). In these cases, we can write $\theta = r(\beta)$ for a function $r : \mathbb{R}^k \to \mathbb{R}^q$ and the estimate of θ is given by

$$\hat{\theta} = r(\hat{\beta})$$

Under Assumption 7.1, we have that $\hat{\theta} \xrightarrow{p} \theta$ if $r(\cdot)$ is continuous at β. This holds by the continuous mapping theorem.

Showing asymptotic normality is somewhat trickier, and I think it is worthwhile to think two distinct cases. First, suppose that $r(\cdot)$ is a linear function; i.e., that we can write $\theta = \mathbf{R}'\beta$ where

$\mathbf{R}$ is a $k \times q$ matrix. In this case, it immediately follows that

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\left(\mathbf{R}'\hat{\beta} - \mathbf{R}'\beta\right) = \mathbf{R}'\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\theta)$$

where

$$\mathbf{V}_\theta = \mathbf{R}'\mathbf{V}_\beta\mathbf{R}$$

Example: Consider the case where $r(\beta) = \beta_1$; this can be alternatively written as $r(\beta) = \mathbf{R}'\beta$ where

$$\mathbf{R} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix}$$

so that $\mathbf{R}$ is a $k \times 1$ vector. Thus,

$$\begin{aligned} \mathbf{V}_\theta &= \begin{pmatrix} 1 & \mathbf{0} \end{pmatrix} \mathbf{V}_\beta \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1k} \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{V}_{11} \end{aligned}$$

i.e., the element in the first row and first column of $\mathbf{V}_\beta$. This explicitly justifies our inference procedures for β_1 discussed above.

Example: Regression Intervals

Suppose that $m(X) := \mathbb{E}[Y|X] = X'\beta$ and that you are interested in constructing a confidence interval for $m(x) := \mathbb{E}[Y|X = x] = x'\beta$ (that is, the value of the conditional CEF at a particular value of the regressors given by x).

The natural way to estimate $m(x)$ is by

$$\hat{m}(x) = x'\hat{\beta}$$

Notice that this can fit into the framework of this section by taking $\theta = m(x)$ so that $\theta = \mathbf{R}'\beta$ for $\mathbf{R} = x$. Thus, notice that

$$\begin{aligned}\sqrt{n}(\hat{m}(x) - m(x)) &= \sqrt{n}(x'\hat{\beta} - x'\beta) \\ &= x'\sqrt{n}(\hat{\beta} - \beta) \\ &\xrightarrow{d} x'\mathcal{N}(0, \mathbf{V}_\beta) = \mathcal{N}(0, x'\mathbf{V}_\beta x)\end{aligned}$$

Thus, we have shown that $\sqrt{n}(\hat{m}(x) - m(x))$ is asymptotically normal with asymptotic variance $x'\mathbf{V}_\beta x)$. We can estimate the asymptotic variance by

$$\hat{\mathbf{V}}_m = x'\hat{\mathbf{V}}_\beta x = x' \left(\frac{1}{n} \sum_{i=1}^n X_i X_i' \right)^{-1} \hat{\mathbf{\Omega}} \left(\frac{1}{n} \sum_{i=1}^n X_i X_i' \right)^{-1} x$$

i.e., we can use exactly the same estimate of $\mathbf{V}_\beta$ that we have been using earlier, just pre-multiplying by x' and post-multiplying by x. Further, notice that $\hat{\mathbf{V}}_m$ is a scalar. Finally, we can construct a 95% confidence interval using essentially the same approach that we used above, that is:

$$\hat{C}_m = \left[x'\hat{\beta} \pm 1.96 \frac{\sqrt{\hat{\mathbf{V}}_m}}{\sqrt{n}} \right]$$

Moreover, if you had some particular $\mathbb{H}_0$ that you wanted to test, you could construct a t-statistic, p-values, etc. along the lines discussed above.

Next, let's move to the case where $r(\cdot)$ is a nonlinear function [as a side-comment, this case generalizes the linear case, so these results cover that case well, but I think it is worth a separate treatment of these two cases]. Under Assumption 7.2, we have that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\theta)$ if $r(\cdot)$ is continuously differentiable in a neighborhood of β and $\mathbf{R} := \nabla r(\beta)$ where $\nabla r(\bar{b}) := \frac{\partial r(b)'}{\partial b}\big|_{b=\bar{b}}$ has rank q. In this case, $\mathbf{V}_\theta = \mathbf{R}'\mathbf{V}_\beta\mathbf{R}$

The above result is just an application of the delta method, but these arguments are important/unfamiliar enough that it is worth explaining in some more detail. Recall that the mean-value theorem says that, if f is a continuous function on $[a, b]$ and differentiable on (a, b), then there exists a $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

In other words, there exists a point in between a and b where the slope of f is equal to the slope of the line connecting $f(a)$ and $f(b)$. Re-arranging implies that

$$f(b) = f(a) + f'(c)(b - a)$$

This is the expressions that will be useful for us (and note that these arguments also go through when f takes a vector argument and/or is vector-valued).

Going back to our case, using a mean-value argument, we can write

$$r(\hat{\beta}) = r(\beta) + \nabla r(\bar{\beta})'(\hat{\beta} - \beta)$$

where $\nabla r(\bar{b}) := \frac{\partial r(b)'}{\partial b}\Big|_{b=\bar{b}}$ (so this is a $k \times q$ dimensional matrix, and plays the role of f' in the mean value theorem above), $\bar{\beta}$ is a vector "between" $\hat{\beta}$ and β (and plays the role of c in the mean value theorem above). Further, notice that by multiplying both sides by $\sqrt{n}$ and re-arranging, it follows that

$$\begin{aligned}
\sqrt{n}(r(\hat{\beta}) - r(\beta)) &= \nabla r(\bar{\beta})'\sqrt{n}(\hat{\beta} - \beta) \\
&= \nabla r(\beta)'\sqrt{n}(\hat{\beta} - \beta) + \Big(\nabla r(\bar{\beta}) - \nabla r(\beta)\Big)'\sqrt{n}(\hat{\beta} - \beta) \\
&= \nabla r(\beta)'\sqrt{n}(\hat{\beta} - \beta) + o_p(1)O_p(1) \\
&= \nabla r(\beta)'\sqrt{n}(\hat{\beta} - \beta) + o_p(1) \\
&\xrightarrow{d} \mathcal{N}(0, \nabla r(\beta)'\mathbf{V}_\beta \nabla r(\beta))
\end{aligned}$$

where the second equality holds by adding and subtracting $\nabla r(\beta)'\sqrt{n}(\hat{\beta} - \beta)$, the third equality holds by the continuous mapping theorem (as long as $\nabla r(b)$ is continuous at β) and because $\bar{\beta}$ is between $\hat{\beta}$ and β, the fourth and fifth equalities hold immediately given the third equality, and the last equality holds because we know the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$ and by the continuous mapping theorem.

Example: Consumer Surplus (H: 7.12)

Problem 7.12 in the textbook concerns running the regression $Y = \alpha + \beta X + e$ where X is a scalar in the case where it is known that $\alpha > 0$ and $\beta < 0$ and then computing the area under the curve defined by the regression line (which is relevant in economics applications for computing consumer surplus) and is given by $A = -\alpha^2/2\beta$. The problem asks to propose an estimator of A and to provide a confidence interval for A. The natural estimator of A is given by

$$\hat{A} = -\frac{\hat{\alpha}^2}{2\hat{\beta}}$$

The key step for coming up with the confidence interval is figuring out the limiting distribution of $\sqrt{n}(\hat{A} - A)$. As a first step, our "usual" arguments for least squares regression imply that

$$\sqrt{n}\begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\beta) \quad \text{where} \quad \mathbf{V}_\beta = \mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1}$$

and $\mathbf{\Omega} = \mathbb{E}[XX'e^2]$ (and where, to keep the expressions from getting too long, I am taking X here to include an intercept, so that $\mathbf{V}_\beta$ is a 2×2 asymptotic variance matrix).

Next, notice that we can write $A = r(\alpha, \beta)$ and $\hat{A} = r(\hat{\alpha}, \hat{\beta})$ where $r(a, b) = -a^2/2b$. This suggests using a delta method type of argument. In particular, using a mean value theorem argument, we can write

$$r(\hat{\alpha}, \hat{\beta}) = r(\alpha, \beta) + \nabla r(\bar{\alpha}, \bar{\beta})' \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} \tag{2}$$

where

$$\nabla r(\bar{a}, \bar{b}) := \begin{bmatrix} \frac{\partial r(a,b)}{\partial a} \\ \\ \frac{\partial r(a,b)}{\partial b} \end{bmatrix}\Bigg|_{a=\bar{a}, b=\bar{b}} = \begin{bmatrix} -\frac{a}{b} \\ \\ \frac{a^2}{2b^2} \end{bmatrix}\Bigg|_{a=\bar{a}, b=\bar{b}}$$

which is the vector of partial derivatives of $r(a, b)$ evaluated at $\bar{a}$ and $\bar{b}$. Plugging this back in to Equation 2 implies that

$$\hat{A} = A + \begin{bmatrix} -\frac{\bar{\alpha}}{\bar{\beta}} \\ \\ \frac{\bar{\alpha}^2}{2\bar{\beta}^2} \end{bmatrix}' \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix}$$

and, by multiplying by $\sqrt{n}$ and adding and subtracting terms, implies that

$$\sqrt{n}(\hat{A} - A) = \begin{bmatrix} -\frac{\alpha}{\beta} \\ \\ \frac{\alpha^2}{2\beta^2} \end{bmatrix}' \sqrt{n}\begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} + \underbrace{\left(\begin{bmatrix} -\frac{\bar{\alpha}}{\bar{\beta}} \\ \\ \frac{\bar{\alpha}^2}{2\bar{\beta}^2} \end{bmatrix} - \begin{bmatrix} -\frac{\alpha}{\beta} \\ \\ \frac{\alpha^2}{2\beta^2} \end{bmatrix}\right)'}_{=o_p(1)} \underbrace{\sqrt{n}\begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix}}_{=O_p(1)} = \begin{bmatrix} -\frac{\alpha}{\beta} \\ \\ \frac{\alpha^2}{2\beta^2} \end{bmatrix}' \sqrt{n}\begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} + o_p(1) \xrightarrow{d} \mathcal{N}(0, V)$$

where the $o_p(1)$ in the first equality arises because (i) $\bar{\alpha}$ is between $\hat{\alpha}$ and α and $\bar{\beta}$ is between $\hat{\beta}$ and β; (ii) $\hat{\alpha} \xrightarrow{p} \alpha$, $\hat{\beta} \xrightarrow{p} \beta$; and (iii) the continuous mapping theorem; and where

$$V = \begin{bmatrix} -\frac{\alpha}{\beta} \\ \\ \frac{\alpha^2}{2\beta^2} \end{bmatrix}' \mathbf{V}_\beta \begin{bmatrix} -\frac{\alpha}{\beta} \\ \\ \frac{\alpha^2}{2\beta^2} \end{bmatrix}$$

Moreover, we can estimate V by

$$\hat{V} = \begin{bmatrix} -\frac{\hat{\alpha}}{\hat{\beta}} \\ \\ \frac{\hat{\alpha}^2}{2\hat{\beta}^2} \end{bmatrix}' \hat{\mathbf{V}}_\beta \begin{bmatrix} -\frac{\hat{\alpha}}{\hat{\beta}} \\ \\ \frac{\hat{\alpha}^2}{2\hat{\beta}^2} \end{bmatrix} \quad \text{where} \quad \hat{\mathbf{V}}_\beta = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2 \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}$$

which is the "usual" estimator of $\mathbf{V}_\beta$. Finally, we can construct a 95

$$\hat{C} = \left[\hat{A} \pm 1.96\frac{\sqrt{\hat{V}}}{\sqrt{n}}\right]$$