These notes come from Chapters 6 and 7 in the textbook and cover the large-sample properties of least squares.

Linear Regression Notes 5: Asymptotic theory for least squares

Review

H: 6.1-6.7

I'll take the concepts of convergence in probability and convergence in distribution as being known (see definitions 6.1 and 6.2 in the textbook)

To start with, we consider the large sample properties (i.e., properties as the sample size gets large) of general estimators, $\hat{\theta}$, of some population parameter θ. The main two properties that we will consider are **consistency** and **asymptotic normality**

Definition:

An estimator $\hat{\theta}$ of θ is **consistent** if $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.

If $\hat{\theta}$ is consistent, this is a guarantee that, given a large enough sample, $\hat{\theta}$ will be "close" to θ.

The key tool for showing that estimators are consistent is the **weak law of large numbers**

Theorem: Weak Law of Large Numbers

If $Y_i \in \mathbb{R}^k$ are iid and $\mathbb{E}||Y|| < \infty$, then as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mathbb{E}[Y]$$

Definition:

An estimator $\hat{\theta}$ of θ is **asymptotically normal** if (for some $\mathbf{V}$)

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}) \quad \text{as } n \to \infty$$

If $\hat{\theta}$ is asymptotically normal, it says that the quantity $\sqrt{n}(\hat{\theta} - \theta)$ should behave like a draw from a normal distribution $\mathcal{N}(0, \mathbf{V})$, given a large enough sample. We will often work towards establishing this sort of result as a key step in conducting statistical inference.

The key tool for showing asymptotic normality is the **central limit theorem**.

Central Limit Theorem:

If $Y_i \in \mathbb{R}^k$ are iid and $\mathbb{E}||Y||^2 < \infty$, then as $n \to \infty$,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}[Y]\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V})$$

where $\mathbf{V} = \text{var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])']$

Let's cover two more tools that are useful for establishing the large sample properties of estimators: the **continuous mapping theorem** and the **delta method**

Continuous Mapping Theorem:

- For convergence in probability: Let $Z_n \in \mathbb{R}^k$ and $g(u) : \mathbb{R}^k \to \mathbb{R}^q$. If $Z_n \xrightarrow{p} c$ as $n \to \infty$ and $g(u)$ is continuous at c, then $g(Z_n) \xrightarrow{p} g(c)$ as $n \to \infty$.
- For convergence in distribution: If $Z_n \xrightarrow{d} Z$ as $n \to \infty$ and $g : \mathbb{R}^k \to \mathbb{R}^q$ has the set of discontinuity points D_g such that $\text{P}(Z \in D_g) = 0$, then $g(Z_n) \xrightarrow{d} g(Z)$ as $n \to \infty$.

These continuous mapping theorems say that continuous functions are limit preserving. Notice that the conditions for the convergence in probability version of the CMT are weaker (they only require g to be continuous at the particular point c) than for the convergence in distribution version (which essentially requires g to be continuous everywhere). The qualification about the set of discontinuity points is a technical one, but comes up enough cases that it is worth including this technical condition.

Delta Method:

Let $\mu \in \mathbb{R}^k$ and $g(u) : \mathbb{R}^k \to \mathbb{R}^q$. If $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \xi$ and $g(u)$ is continuously differentiable in a neighborhood of μ, then as $n \to \infty$,

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{d} \mathbf{G}'\xi$$

where $\mathbf{G}(u) = \frac{\partial g(u)'}{\partial u}$ and $\mathbf{G} = \mathbf{G}(\mu)$. As a leading example, if $\xi \sim \mathcal{N}(0, \mathbf{V})$, then as $n \to \infty$,

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, \mathbf{G}'\mathbf{V}\mathbf{G})$$

Stochastic Order Symbols:

It will be helpful to sometimes have a notation for random variables that converge in probability to zero or are stochastically bounded. We write

$$Z_n = o_p(1)$$

to mean that $Z_n \xrightarrow{p} 0$ as $n \to \infty$. And we write

$$Z_n = O_p(1)$$

to indicate that Z_n is "bounded in probability" – you can see the textbook for a formal definition, but you should take this to mean that Z_n does not diverge to positive or negative infinity as $n \to \infty$. The textbook provides a number of properties of $o_p(1)$ and $O_p(1)$. I think the most useful ones are that

$$O_p(1) + o_p(1) = O_p(1) \qquad O_p(1)o_p(1) = o_p(1)$$

which say that (i) if you add something that is bounded in probability to something that converges to 0 then the result will be bounded in probability, and (ii) that if you multiply something bounded in probability to something that converges in probability to 0 then the result will converge in probability to 0. These are implications of the continuous mapping theorem.

Asymptotic Theory for Least Squares

The asymptotic theory for least squares applies both to linear projection model and to the linear CEF model. Therefore, in this section, we only use the weaker assumptions of the linear projection model. That is, we use the following assumptions throughout this section

Assumption 7.1

1. The variables $\{(Y_i, X_i)\}_{i=1}^n$ are iid
2. $\mathbb{E}[Y^2] < \infty$
3. $\mathbb{E}||X||^2 < \infty$
4. $\mathbb{E}[XX']$ is positive definite

Consistency of Least Squares Estimator

H: 7.2

Step 1: Weak Law of Large Numbers. Recall that

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

Next, notice that

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

which holds by the weak law of large numbers (which requires the iid assumption and that $\mathbb{E}[XX'] < \infty$ and $\mathbb{E}[XY] < \infty$, both of which hold by Assumption 7.1)

Step 2: Continuous Mapping Theorem. Next, notice that, we can write

$$\hat{\beta} = g(\hat{\mathbb{E}}[XX'], \hat{\mathbb{E}}[XY])$$

where $g(\mathbf{A}, b) = \mathbf{A}^{-1}b$. This is a continuous function of $\mathbf{A}$ and b at all values of the arguments such that $\mathbf{A}^{-1}$ exists. Assumption 7.1 includes that $\mathbb{E}[XX']$ is positive definite which implies that $\mathbb{E}[XX']^{-1}$ exists. Thus, $g(\mathbf{A}, b)$ is continuous at $\mathbf{A} = \mathbb{E}[XX']$ and we can apply the "convergence in probability" version of the CMT; that is,

$$\begin{aligned}\hat{\beta} &\xrightarrow{p} g(\mathbb{E}[XX'], \mathbb{E}XY) \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY] = \beta\end{aligned}$$

Asymptotic Normality

H: 7.3

For this section, we strengthen Assumption 7.1.

Assumption 7.2

In addition to Assumption 7.1

1. $\mathbb{E}[Y^4] < \infty$
2. $\mathbb{E}||X||^4 < \infty$

Next, we will establish the limiting distribution of $\hat{\beta}$. Plugging $Y_i = X_i'\beta + e_i$ into Equation 1 implies that

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

Multiplying by $\sqrt{n}$ and re-arranging implies that

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_ie_i \tag{2}$$

Step 1: Central Limit Theorem. First, notice that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_ie_i\xrightarrow{d}\mathcal{N}(0,\mathbf{\Omega})$$

where $\mathbf{\Omega}=\mathbb{E}[Xe(Xe)']=\mathbb{E}[XX'e^2]$.

Let's explain carefully why the central limit theorem applies here. First, we have that (Y_i,X_i) are iid, which implies that any function of (Y_i,X_i) is also iid (and this includes $e_i=Y_i-X_i'\beta$ and X_ie_i). Also, notice that $\mathbb{E}[Xe]=0$ so that the inside term of the summation above has mean 0. Finally, to invoke the central limit theorem, we need to show that the second moments of Xe exist; i.e., that $\mathbb{E}||Xe||^2<\infty$. This argument is technical, but holds under Assumption 7.2; if you are interested in this, see the Technical Details section below (you will see where we use the fourth moment conditions in Assumption 7.2). That $\mathbf{\Omega}=\text{var}(Xe)=\mathbb{E}[XX'e^2]$ is a direct consequence of applying the central limit theorem.

Technical Details: Here we show that the second moments of Xe exist under Assumption 7.2. Below, we use the following inequalities:

Minkowski's Inequality: For $p \geq 1$, $(\mathbb{E}||X+Y||^p)^{1/p} \leq (\mathbb{E}||X||^p)^{1/p} + (\mathbb{E}||Y||^p)^{1/p}$
Schwarz Inequality: $|a'b| \leq ||a||\ ||b||$

As a first step, let's show that Assumption 7.2 implies that $\mathbb{E}[e^4] < \infty$. Notice that,

$$\begin{aligned}
\mathbb{E}[e^4]^{1/4} &= \mathbb{E}[(Y - X'\beta)^4]^{1/4} \\
&\leq \mathbb{E}[Y^4]^{1/4} + \mathbb{E}[(X'\beta)^4]^{1/4} \\
&\leq \mathbb{E}[Y^4]^{1/4} + (\mathbb{E}||X||^4)^{1/4}||\beta|| \\
&< \infty
\end{aligned}$$

where the second line uses Minkowski's inequality, the third inequality holds by the Schwarz inequality (to be clear on this part, notice that $\mathbb{E}[(X'\beta)^4]^{1/4} = \mathbb{E}[|X'\beta|^4]^{1/4} \leq \mathbb{E}[(||X||\ ||\beta||)^4]^{1/4} = \mathbb{E}[||X||^4||\beta||^4]^{1/4} = \mathbb{E}[||X||^4]^{1/4}||\beta|| < \infty$), and the last inequality holds by Assumption 7.2. That $\mathbb{E}[e^4]^{1/4} < \infty$ implies that $\mathbb{E}[e^4] < \infty$. To show the main result, we will also use the following two inequalities:

Expectation Inequality: For a random vector $Y \in \mathbb{R}^m$ with $\mathbb{E}||Y|| < \infty$, $||\mathbb{E}[Y]|| \leq \mathbb{E}||Y||$.
Cauchy-Schwarz Inequality: $\mathbb{E}||X'Y|| \leq (\mathbb{E}||X||^2)^{1/2}(\mathbb{E}||Y||^2)^{1/2}$

To show that the second moments of Xe exist, we will just directly show that all the elements of the second moment matrix, $\mathbf{\Omega}$, are finite. In particular, generically consider the (j,l) element of $\mathbf{\Omega}$ which is given by $\mathbb{E}[X_j X_l e^2]$ (we want to show that this is finite). Therefore, consider

$$\begin{aligned}
|\mathbb{E}[X_j X_l e^2]| &\leq \mathbb{E}|X_j X_l e^2| \\
&= \mathbb{E}[|X_j|\ |X_l|e^2] \\
&\leq \mathbb{E}[X_j^2 X_l^2]^{1/2}\mathbb{E}[e^4]^{1/2} \\
&\leq \left(\mathbb{E}[X_j^4]^{1/2}\mathbb{E}[X_l^4]^{1/2}\right)^{1/2}\mathbb{E}[e^4]^{1/2} \\
&= \mathbb{E}[X_j^4]^{1/4}\mathbb{E}[X_l^4]^{1/4}\mathbb{E}[e^4]^{1/2} \\
&< \infty
\end{aligned}$$

where the first equality holds by the expectation inequality, the second equality holds because of the absolute value, the third equality holds by the Cauchy-Schwarz inequality, the fourth equality holds by applying the Cauchy-Schwarz inequality again, the fifth equality holds immediately, and the last equality holds by Assumption 7.2 and because $\mathbb{E}[e^4] < \infty$ (which we showed right before).

Combining this with Equation 2, we have that

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

where $\mathbf{V}_\beta = \mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1}$ and which holds by the continuous mapping theorem.

$\mathbf{V}_\beta$ is called the **asymptotic variance matrix** of $\hat{\beta}$. $\mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1}$ is called a "sandwich form". It is called this because $\mathbf{\Omega}$ is sandwiched by $\mathbb{E}[XX']^{-1}$ (sometimes $\mathbf{\Omega}$ is called the "meat" and $\mathbb{E}[XX']^{-1}$ is called the "bread"). Many asymptotic variance matrices have a similar form.

The previous result is the basis for hypothesis testing/inference, constructing confidence intervals, etc. To operationalize it, though, we need to construct an estimator of $\mathbf{V}_\beta$. Before doing that, let's introduce one relatively common simplification.

Homoskedasticity Assumption:

$\mathbb{E}[e^2|X] = \sigma^2$.

Homoskedasticity says that the second moment of the error term does not vary across different values of X. This is often constrasted with **heteroskedasticity** which amounts to just not making the homoskedasticity assumption. Most applications in economics do not invoke the homoskedasticity assumption mainly because, often, we do not "need" it. That said, as we will see below, it is useful for simplifying some expressions and serves as a useful benchmark in many cases.

Notice that, under homoskedasticity, we can simplify the expression for $\mathbf{\Omega}$ (I use the notation $\mathbf{\Omega}_0$ to indicate that this is the expression for $\mathbf{\Omega}$ under homoskedasticity):

$$\mathbf{\Omega}_0 = \mathbb{E}\Big[XX' \underbrace{\mathbb{E}[e^2|X]}_{\sigma^2}\Big] = \sigma^2\mathbb{E}[XX']$$

where the first equality holds by the law of iterated expectations, and the second equality holds by homoskedasticity. Plugging this back in to the expression for $\mathbf{V}_\beta$, it will also simplify (again, I switch the notation to indicate the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$ under homoskedasticity):

$$\mathbf{V}_0 = \sigma^2\mathbb{E}[XX']^{-1}$$

which holds by plugging in $\mathbf{\Omega}_0$ into the expression for $\mathbf{V}_\beta$ and cancelling.

Discussion:

What we have shown is that the sampling distribution of quantity $\sqrt{n}(\hat{\beta} - \beta)$ is $\mathcal{N}(0, \mathbf{V}_\beta)$, as long as we have a large sample. In practice, we only have one "draw" from this distribution which corresponds to the sample that we actually have (and, here, we are ignoring that we do not know the value of the population parameter β). What we have shown is that (given a large enough sample) this "draw" should amount to a draw from $\mathcal{N}(0, \mathbf{V}_\beta)$ – we will exploit this heavily when we discuss inference soon.

One related question is how large the sample needs to be for our results on consistency and

asymptotic normality of $\hat{\beta}$ to hold. Although you may have heard various rules-of-thumb, my sense is that there is no general rule here. In particular, one can come up with cases where it would take an extremely large number of observations before the asymptotic approximation would work very well (see p.167 for an example). That said, most work in economics uses at least hundreds of observations. Estimating more complicated models may tend to require more observations for these approximations to work well.

Consistency of Error Variance Estimators

H: 7.5

As discussed above, in order to use the asymptotic normality result above, we need to consistently estimate $\mathbf{V}_\beta$. In this section, we consider the simpler case of estimating $\mathbf{V}_0$ (i.e., the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$ under homoskedasticity. Notice that, by the continuous mapping theorem, we can consistently estimate $\mathbf{V}_0$ by consistently estimating σ^2 and $\mathbb{E}[XX']$. Estimating $\mathbb{E}[XX']$ is straightforward – we can just use the analogy principle. On the other hand, estimating $\sigma^2 = \mathbb{E}[e^2]$ is conceptually more challenging, and we consider this next. Using the analogy principle would suggest estimating σ^2 by

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2$$

but this estimator is infeasible since we do not observe e_i. Instead, let's consider the estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

where $\hat{e}_i$ is the **residual** that is defined as

$$\hat{e}_i = Y_i - X_i'\hat{\beta}$$

which is the difference between the actual outcome and $X_i\hat{\beta}$ (the fitted value from the regression). Notice that, $\hat{e}_i$ is something that we can actually recover because it depends on the estimated $\hat{\beta}$ rather than, say, the population parameter β. Notice that, by plugging in $Y_i = X_i'\beta$ into the expression for $\hat{e}_i$, we have that

$$\begin{aligned}
\hat{e}_i &= Y_i - X_i'\hat{\beta} \\
&= X_i'\beta + e_i - X_i'\hat{\beta} \\
&= e_i - X_i'(\hat{\beta} - \beta)
\end{aligned}$$

which implies that

$$\hat{e}_i^2 = e_i^2 - 2e_iX_i'(\hat{\beta} - \beta) + (\hat{\beta} - \beta)'X_iX_i'(\hat{\beta} - \beta)$$

so that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n}e_iX_i'\right)(\hat{\beta}-\beta) + (\hat{\beta}-\beta)\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)(\hat{\beta}-\beta)$$

Then, since,

$$\frac{1}{n}\sum_{i=1}^{n}e_i^2 \xrightarrow{p} \mathbb{E}[e^2] = \sigma^2$$
$$\hat{\beta}-\beta \xrightarrow{p} 0$$
$$\frac{1}{n}\sum_{i=1}^{n}X_iX_i' \xrightarrow{p} \mathbb{E}[XX']$$

it follows by the continuous mapping theorem that

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2$$

Moreover, this implies that

$$\hat{\mathbf{V}}_0 := \hat{\sigma}^2\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)^{-1} \xrightarrow{p} \sigma^2\mathbb{E}[XX]^{-1} = \mathbf{V}_0$$

so that $\hat{\mathbf{V}}_0$ is consistent for $\mathbf{V}_0$.

Heteroskedastic Covariance Matrix Estimation

H: 7.7

Next, we consider estimating $\mathbf{V}_\beta$. The natural estimator is

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

where $\hat{\mathbf{\Omega}}$ is an estimate of $\mathbf{\Omega}$ given by

$$\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}X_iX_i'\hat{e}_i^2$$

We aim to show that $\hat{\mathbf{\Omega}}$ is consistent for $\mathbf{\Omega}$. To this end, notice that

$$\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}X_iX_i'e_i^2 + \frac{1}{n}\sum_{i=1}^{n}X_iX_i'(\hat{e}_i^2 - e_i^2)$$

which holds by adding and subtracting terms. Then, notice that

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 \xrightarrow{p} \mathbb{E}[XX'e^2] = \mathbf{\Omega}$$

It remains to show be shown that

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' (\hat{e}_i^2 - e_i^2) \xrightarrow{p} 0 \tag{3}$$

Given our earlier result on $\hat{\sigma}^2$ being consistent for σ^2, it is perhaps not surprising that this term converges to 0 though the arguments are more challenging (if you are interested, please check out the Technical Details box below).

Technical Details: Here I show that the claim in Equation 3 is true. To start with, let me briefly introduce some useful concepts related to matrix norms and useful inequalities for matrix norms. Below, $\mathbf{A}$ and $\mathbf{B}$ are notation for matrices.

Frobenius/Matrix Norm $||\mathbf{A}|| = ||\text{vec}(\mathbf{A})||$
Schwarz Inequality $||\mathbf{AB}|| \leq ||\mathbf{A}||\,||\mathbf{B}||$
Triangle Inequality: $||\mathbf{A}+\mathbf{B}|| \leq ||\mathbf{A}|| + ||\mathbf{B}||$
Holder's Inequality: For any $p > 1$ and $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, $\mathbb{E}||X'Y|| \leq \left(\mathbb{E}||X||^p\right)^{1/p}\left(\mathbb{E}||Y||^q\right)^{1/q}$

The Frobenius norm is a matrix norm (there are others) that "converts" the matrix into a vector and then applies the Euclidean norm to that vector. The next two inequalities say that versions of the Schwarz and triangle inequalities apply to matrices. Next, notice that

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i'(\hat{e}_i^2 - e_i^2)\right\| &\leq \frac{1}{n}\sum_{i=1}^{n}\left\|X_i X_i'(\hat{e}_i^2 - e_i^2)\right\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}||X_i||^2|\hat{e}_i^2 - e_i^2| \qquad (4)\end{aligned}$$

where the first inequality holds by the triangle inequality and the second inequality holds by applying the Schwarz inequality twice. Now consider

$$\begin{aligned}|\hat{e}_i^2 - e_i^2| &= |-2e_i X_i'(\hat{\beta}-\beta) + (\hat{\beta}-\beta)'X_i X_i'(\hat{\beta}-\beta)| \\ &\leq 2|e_i X_i'(\hat{\beta}-\beta)| + (\hat{\beta}-\beta)'X_i X_i'(\hat{\beta}-\beta) \\ &= 2|e_i|\;|X_i'(\hat{\beta}-\beta)| + |(\hat{\beta}-\beta)'X_i|^2 \\ &\leq 2|e_i|\;||X_i||\;||\hat{\beta}-\beta|| + ||X_i||^2\;||\hat{\beta}-\beta||^2\end{aligned}$$

where the first equality holds by plugging in from above the difference between $\hat{e}_i^2$ and e_i^2, the second inequality holds by the triangle inequality (the second term is positive because it is quadratic), the third equality holds by properties of absolute value, the fourth inequality holds by the Schwarz inequality. Using this expression back in Equation 4 implies that

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i'(\hat{e}_i^2 - e_i^2)\right\| \leq 2\left(\frac{1}{n}\sum_{i=1}^{n}||X_i||^3\;|e_i|\right)\;||\hat{\beta}-\beta|| + \frac{1}{n}\sum_{i=1}^{n}||X_i||^4\;||\hat{\beta}-\beta||^2$$

The second term converges to 0 because $n^{-1}\sum_{i=1}^{n}||X_i||^4 \xrightarrow{p} \mathbb{E}[X^4]$ and because $||\hat{\beta}-\beta|| \xrightarrow{p} 0$. For the first term $||\hat{\beta}-\beta|| \xrightarrow{p} 0$, and then consider

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}||X_i||^3\;|e_i| &\xrightarrow{p} \mathbb{E}[||X||^3\;|e|] \\ &\leq \mathbb{E}\left[(||X||^3)^{4/3}\right]^{3/4}\mathbb{E}[e^4]^{1/4} \\ &= \mathbb{E}[||X||^4]^{3/4}\mathbb{E}[e^4]^{1/4} \\ &< \infty\end{aligned}$$

where the first equality holds by the weak law of large numbers, the second equality holds using Holder's inequality (using $||X||^3$ and $|e|$ and setting $p = 4/3$ and $q = 4$), the third equality by canceling the inside exponents, and the last inequality by Assumption 7.2 and that we showed that $\mathbb{E}[e^4] < \infty$ in the previous Technical Details box.

Inference

Asymptotic Standard Errors

H: 7.11-7.13, 7.16, 9.7, 9.9

Inference and hypothesis testing were covered in detail in 8070. This section provides a brief review along with (brief) explanations in the context of regression.

Next, let us return to our results on asymptotic normality of $\hat{\beta}$ in order to see how this is useful for hypothesis testing.

We define the **standard error** of the $\hat{\beta}_j$ (the jth element of $\hat{\beta}$) as

$$\text{se}(\hat{\beta}_j) = \frac{\sqrt{\hat{\mathbf{V}}_{\beta,jj}}}{\sqrt{n}}$$

where $\hat{\mathbf{V}}_{\beta,jj}$ is the (j,j) element of $\hat{\mathbf{V}}_\beta$. As a side-comment, I define this slightly differently from the book. I use $\hat{\mathbf{V}}_\beta$ (the asymptotic variance matrix) rather than $\mathbf{V}_{\hat{\beta}} = \text{var}(\hat{\beta}|\mathbf{X})$. These are different from each other by a factor of $\sqrt{n}$; in particular, $\hat{\mathbf{V}}_\beta$ does not go to 0 as $n \to \infty$.

It is common in applications in economics to report $\hat{\beta}$ along with standard errors for each estimated parameter.

In many cases, a researcher may be interested in testing a hypothesis for a single parameter in a regression. We have typically made the regressor of interest X_1, so I'll follow that convention here, but analogous arguments apply for other regressors. Consider $\mathbb{H}_0 : \beta_1 = \theta_0$ (where θ_0 is a hypothesized value for β_1; this implies that θ_0 is known to us, and, e.g., by far the most common choice is $\theta_0 = 0$), then the **t-statistic** is given by

$$t = \frac{\hat{\beta}_1 - \theta_0}{\text{s.e.}(\hat{\beta}_1)}$$

Recall that, if $\mathbb{H}_0$ is true, then t will behave like a draw from a standard normal distribution. On the other hand, if $\mathbb{H}_0$ is false, then t will diverge (i.e., go to positive or negative infinity) as $n \to \infty$. This difference is the basis for us to conduct inference. In particular, when $|t| > c$ (where c is some critical value such as 1.96 when the significance level is 5%), then one would reject $\mathbb{H}_0$ and otherwise fail to reject $\mathbb{H}_0$.

The approach to inference discussed so far has been to compute a t-statistic and then to make a binary decision to either reject or fail to reject $\mathbb{H}_0$. This approach has some inherent issues. The textbook gives the example of a t-statistic equal to 1.7 relative to one that is equal to 2.0. Given a 5% significance level, these t-statistics lead to different decisions. However, it is immediately clear that the strength of evidence against $\mathbb{H}_0$ is not really much different between these two cases.

An alternative approach is to report an (asympotic) **p-value**. The p-value is the probability of getting a test-statistic as large (in absolute value) as we did given that $\mathbb{H}_0$ is true (an alternative interpretation is that p is the smallest value of α for which the test would reject $\mathbb{H}_0$). Given that

$t \xrightarrow{d} Z \sim \mathcal{N}(0,1)$,

$$\begin{aligned} p &= \mathrm{P}(Z < -|t|) + \mathrm{P}(Z > |t|) \\ &= 2(1 - \Phi(|t|)) \end{aligned}$$

where the second line follows by symmetry of Z. Unlike the binary decision rule that we have discussed previously, the p-value provides continuous information. For example, if we calculate that $p = 0.06$, we would not reject $\mathbb{H}_0$ at the 5% significance level, but getting a t-statistic this large in absolute value is still relatively uncommon if $\mathbb{H}_0$ is true. Similarly, if $p = 0.00001$, this would indicate very strong evidence against $\mathbb{H}_0$.

Next, we also have enough information to compute a **confidence interval**. The most common version of a confidence interval is the one given by

$$\widehat{C} = [\hat{\theta} - c_{1-\alpha/2}\text{s.e.}(\hat{\theta}), \hat{\theta} + c_{1-\alpha/2}\text{s.e.}(\hat{\theta})]$$

where, for example, if $\alpha = 0.05$, then $c_{1-\alpha/2} = c_{.975} = 1.96$ (because 1.96 is the 97.5th percentile of a standard normal distribution).

Finally, in some cases, a researcher may be interested in testing multiple hypotheses. For example, one might be interested in testing $\mathbb{H}_0 : \beta_1 = \beta_2 = 0$. More generally, consider $\theta = r(\beta)$ where $r : \mathbb{R}^k \to \mathbb{R}^q$ and $\hat{\theta} = r(\hat{\beta})$. Further, suppose that we know $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\theta)$ for some $\mathbf{V}_\theta$ (it require some work to show this sort of result, e.g., see the discussion in the next section, but you can also fit in simple examples here such as $\theta = \beta$). And suppose that we are interested in $\mathbb{H}_0 : \theta = \theta_0$. It is hard to operationalize the t-statistic that we talked about above. Instead, we will consider a Wald statistic:

$$W = n(\hat{\theta} - \theta_0)' \hat{\mathbf{V}}_\theta^{-1} (\hat{\theta} - \theta_0)$$

Notice that this is a number that we can compute (given a value of θ_0) and that it is a scalar. As for t above, let's consider the behavior of W under $\mathbb{H}_0$ and under $\mathbb{H}_1 : \theta \neq \theta_0$. When $\mathbb{H}_0$ is true, then one can show that W will behave like a draw from a χ^2_q distribution (i.e., a chi-squared distribution with q degrees of freedom). On the other hand, if $\mathbb{H}_0$ is not true, then W will diverge to infinity as $n \to \infty$.

As for the t-statistic, this different behavior under $\mathbb{H}_0$ relative to $\mathbb{H}_1$ provides an approach to inference. For testing multiple restrictions like this, I think that it is most common to report a p-value. Here, you can calculate a p-value by

$$\text{p-value} = 1 - G_q(W)$$

where G_q is the cdf of a chi-square random variable with q degrees of freedom.

Because the distribution of W under $\mathbb{H}_0$ depends on the degrees of freedom q (i.e., the number of restrictions being tested), it is harder to "just remember" critical values. That said, it is easy to compute the p-value above in `R` using the function `pchisq` (which computes the cdf of a chi-square random variable). For example, suppose that you calculate $W = 7$ and that $q = 2$. Then, the p-value can be calculated as

```
p <- 1 - pchisq(7,df=2)
round(p,4)
```

```
[1] 0.0302
```

so that the p-value is about 0.03 (indicating that you would reject $\mathbb{H}_0$ at the 5% significance level).

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 $\hat{A} = r(\alpha, \beta)$ 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

$$r(\hat{\alpha}, \hat{\beta}) = r(\alpha, \beta) + \nabla r(\bar{\alpha}, \bar{\beta})' \begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix} \tag{5}$$

where

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

which is the vector of partial derivatives of $r(a, b)$ evaluated at $\bar{a}$ and $\bar{b}$. Plugging this back in to Equation 5 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

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