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

Linear Regression Notes 4: 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 $\mathbf{V}$)

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

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 $N(0, \mathbf{V})$, 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 $n \to \infty$,

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

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

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

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

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i \tag{2}$$

Step 1: Central Limit Theorem. First, notice that

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

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 $X_i e_i$). Also, notice that $\mathbb{E}[Xe] = 0$ which implies that $\text{var}(Xe) = \Omega$. Finally, to invoke the central limit theorem, we need to show that our assumptions imply that all of the elements of Ω are finite (if you are interested in this, see the technical details section below).

Technical Details: As a step in this direction, let's first show that Assumption 7.2 implies that $\mathbb{E}[e^4] < \infty$.

Minkowski's Inequality: $(\mathbb{E}||X+Y||^p)^{1/p} \leq (\mathbb{E}||X||^p)^{1/p} + (\mathbb{E}||Y||^p)^{1/p}$

Schwarz Inequality: $|a'b| \leq ||a||\ ||b||$

Therefore,

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

where the second line uses Minkowski's inequality, the third inequality holds by the Schwarz inequality; to be clear on this part, notice that $\mathbb{E}[(X'\beta)^4]^{1/4} = \mathbb{E}[|X'\beta|^4]^{1/4} \leq \mathbb{E}[(||X||\ ||\beta||)^4]^{1/4} = \mathbb{E}[||X||^4||\beta||^4]^{1/4} = \mathbb{E}[||X||^4]^{1/4}||\beta|| < \infty$. That $\mathbb{E}[e^4]^{1/4} < \infty$ implies that $\mathbb{E}[e^4] < \infty$.

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: $\mathbb{E}||X'Y|| \leq (\mathbb{E}||X||^2)^{1/2}(\mathbb{E}||Y||^2)^{1/2}$

Next, the (j,l) element of $\mathbf{\Omega}$ 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

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

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.

Discussion: What we have shown is that the sampling distribution of quantity $\sqrt{n}(\hat{\beta} - \beta)$ is $N(0, \mathbf{V}_\beta)$, 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 $N(0, \mathbf{V}_\beta)$ – 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

Next, we consider estimating $\sigma^2 = \mathbb{E}[e^2]$. Let's consider the estimator

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

Notice 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

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

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

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

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

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

It remains to show be shown that

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

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, see the technical details below).

Technical Details: To start with, recall a matrix version of the triangle inequality:

Triangle Inequality: $||\mathbf{A}+\mathbf{B}|| \leq ||\mathbf{A}|| + ||\mathbf{B}||$

Next, notice that

$$\begin{aligned}\left|\left|\frac{1}{n}\sum_{i=1}^{n} X_i X_i'(\hat{e}_i^2 - e_i^2)\right|\right| &\leq \frac{1}{n}\sum_{i=1}^{n} \left|\left|X_i X_i'(\hat{e}_i^2 - e_i^2)\right|\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} ||X_i||^2 |\hat{e}_i^2 - e_i^2| \end{aligned} \tag{3}$$

where the first inequality holds by the triangle inequality and the second inequality holds by the Schwarz inequality. 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 3 implies that

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

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 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 $\mathbb{E}[e^4] < \infty$.

Thus, we have shown that $\hat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$. It immediately follows from the weak law of large numbers and the continuous mapping theorem that $\hat{\mathbf{V}}_\beta \xrightarrow{p} \mathbb{E}[XX']^{-1}\mathbf{\Omega}\mathbb{E}[XX']^{-1} = \mathbf{V}_\beta$. This implies that $\hat{\mathbf{V}}_\beta$ is consistent for $\mathbf{V}_\beta$.

> **Practice:** Relative to the above discussion, suppose that homoskedasticity also holds; that is, $\mathbb{E}[e^2|X] = \sigma^2$. How does $\mathbf{V}_\beta$ simplify in this case? Given this simplification, propose an estimator of $\mathbf{V}_\beta$ and show that it is consistent.

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.

Under Assumption 7.2, we have that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V}_\theta)$ if $r(\cdot)$ is continuously differentiable in a neighborhood of β and $\mathbf{R} := \frac{\partial}{\partial \beta} r(\beta)'$ has rank q. In this case, $\mathbf{V}_\theta = \mathbf{R}'\mathbf{V}_\beta\mathbf{R}$

> **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. In this case,
>
> $$\begin{aligned} \mathbf{V}_\theta &= \begin{pmatrix} 1 & \mathbf{0} \end{pmatrix} \mathbf{V}_\beta \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1k} \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{V}_{11} \end{aligned}$$
>
> i.e., the element in the first row and first column of $\mathbf{V}_\beta$.