LECTURE NOTE 8

LIMIT DISTRIBUTION FOR THE LEAST-SQUARES ESTIMATOR

1. CONSISTENT ESTIMATION OF LINEAR PREDICTORS

We start by showing that the least-squares estimator converges, in a certain sense, to the linear predictor coefficients. We observe the realizations of the random variables $Y_i, X_{i1}, \ldots, X_{iK}$ for $i = 1, \ldots, n$. Let $X_i' = (\, X_{i1} \quad \ldots \quad X_{iK}\,)$. As in Note 6, assume random sampling, so that the (Y_i, X_i) are independent and identically distributed (i.i.d.) from some joint distribution. Our notation for the linear predictor is

$$E^*(Y_i \,|\, X_i) = X_i'\beta,$$

and the least-squares estimator is

$$b = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i.$$

The sense in which b converges to β is *convergence in probability*:

Definition. The sequence of random variables Q_n converges in probability to a constant α if

$$\lim_{n\to\infty} \text{Prob}(|Q_n - \alpha| > \epsilon) = 0$$

for all $\epsilon > 0$. Notation: $Q_n \xrightarrow{p} \alpha$.

We have frequently used the intuitive argument that sample moments can be used as estimators for population moments. This is justified in large samples by the law of large numbers:

Law of Large Numbers. If W_i i.i.d. and $E(|W_i|) < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{p} E(W_1).$$

It is convenient to work with convergence in probability because it interacts nicely with continuous functions:

Slutsky Theorem. (i) If the sequence of random variables Q_n takes on values in $\mathcal{R}^J$, $Q_n \xrightarrow{p} \alpha$, and the function $g\colon \mathcal{R}^J \to \mathcal{R}^M$ is continuous at α, then

$$g(Q_n) \xrightarrow{p} g(\alpha).$$

We say that b is a *consistent estimator* of β if $b \xrightarrow{p} \beta$.

Claim 1. $b \xrightarrow{p} \beta$ as $n \to \infty$.

Proof. By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{p} E(X_1 Y_1),$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} E(X_1 X_1').$$

Since b is a continuous function of these sample moments, Slutsky's theorem implies that

$$b \xrightarrow{p} [E(X_1 X_1')]^{-1} E(X_1 Y_1) = \beta. \quad \diamond$$

2. LIMIT DISTRIBUTION

Define the prediction error

$$U_i = Y_i - E^*(Y_i \mid X_i),$$

so that

$$Y_i = X_i'\beta + U_i, \quad E(X_i U_i) = 0.$$

Substitute this expression for Y_i into the formula for the least-squares estimator:

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

Now look at $\sqrt{n}(b-\beta)$:

$$
\sqrt{n}(b-\beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i U_i.
$$

Because b is converging to β, its distribution is becoming degenerate, with the probability piling up in a shrinking neighborhood of β. So we multiply by $\sqrt{n}$ to obtain a nondegenerate limit distribution. Define $G_i = X_i U_i$. Then G_i is i.i.d. (since it is a function of (Y_i, X_i)), $E(G_i) = 0$, and

$$
\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(G_i) = \mathrm{Cov}(G_1).
$$

So $\sqrt{n}$ is the right factor to stabilize the variance. Now we can appeal to the central limit theorem to obtain a normal distribution.

Let W be a $K \times 1$ random variable with a a $\mathcal{N}(0,\Sigma)$ distribution. A sequence of random variables S_n converges in distribution to $\mathcal{N}(0,\Sigma)$ if for any (well-behaved) subset A of $\mathcal{R}^K$, we have

$$
\lim_{n\to\infty} \mathrm{Prob}(S_n \in A) = \mathrm{Prob}(W \in A).
$$

Notation: $S_n \xrightarrow{d} \mathcal{N}(0,\Sigma)$.

Central Limit Theorem. If the $K \times 1$ random variables G_i are independent and identically distributed with $E(G_i) = 0$ and $\mathrm{Cov}(G_i) = \Sigma$, then

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_i \xrightarrow{d} \mathcal{N}(0,\Sigma).
$$

There is a second part to the Slutsky theorem:

Slutsky Theorem. (ii) Let S_n be a sequence of $K \times 1$ random variables with $S_n \xrightarrow{d} \mathcal{N}(0, \Sigma)$, and let Q_n be a sequence of $J \times K$ random variables with $Q_n \xrightarrow{p} \alpha$, a constant. Then

$$Q_n S_n \xrightarrow{d} \alpha \cdot \mathcal{N}(0, \Sigma) = \mathcal{N}(0, \alpha \Sigma \alpha').$$

Now we can use the law of large numbers, the central limit theorem, and the Slutsky theorem to obtain a limit distribution for the least-squares estimator.

Claim 2. $\sqrt{n}(b - \beta) \xrightarrow{d} \mathcal{N}(0, \alpha \Sigma \alpha')$, where

$$\alpha = [E(X_1 X_1')]^{-1}, \quad \Sigma = E(U_1^2 X_1 X_1').$$

Proof. Define $G_i = X_i U_i$ and note that $E(G_i) = 0$, $\mathrm{Cov}(G_i) = E(G_i G_i') = \Sigma$. By the central limit theorem,

$$S_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} G_i \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

By the law of large numbers and Slutsky (i),

$$Q_n \equiv \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \xrightarrow{p} [E(X_1 X_1')]^{-1} = \alpha.$$

Then by Slutsky (ii),

$$\sqrt{n}(b - \beta) = Q_n S_n \xrightarrow{d} \alpha \mathcal{N}(0, \Sigma) = \mathcal{N}(0, \alpha \Sigma \alpha'). \quad \diamond$$

3. CONFIDENCE INTERVAL

Define $\Lambda = \alpha \Sigma \alpha'$, so that Claim 2 gives

$$\sqrt{n}(b - \beta) \xrightarrow{d} \mathcal{N}(0, \Lambda).$$

(Λ is capital lambda.) Λ is the covariance matrix for the limit distribution and is known as the asymptotic covariance matrix. We can use this limit distribution to construct a confidence interval. In order to obtain a consistent estimate of Λ, let $\hat{U}_i = Y_i - X_i'b$ and define

$$\hat{\Lambda} = \hat{\alpha}\hat{\Sigma}\hat{\alpha}',$$

with

$$\hat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2 X_i X_i'.$$

We can use the law of large numbers and Slutsky (i) to show that

$$\hat{\Lambda} \xrightarrow{p} \Lambda.$$

As in Section 4 of Note 7, we shall obtain a confidence interval for a linear combination of the coefficients:

$$l'\beta = \sum_{j=1}^{K} l_j \beta_j.$$

Slutsky (i) implies that

$$(l'\hat{\Lambda}l)^{1/2} \xrightarrow{p} (l'\Lambda l)^{1/2},$$

and Slutsky (ii) implies that

$$\frac{l'[\sqrt{n}(b-\beta)]}{(l'\hat{\Lambda}l)^{1/2}} \xrightarrow{d} \frac{1}{(l'\Lambda l)^{1/2}}\mathcal{N}(0, l'\Lambda l) = \mathcal{N}(0,1).$$

Define the standard error as

$$\text{SE} = (l'\hat{\Lambda}l/n)^{1/2}.$$

We have established

Claim 3.

$$\frac{l'(b-\beta)}{\text{SE}} \xrightarrow{d} \mathcal{N}(0,1).$$

The ratio $l'(b-\beta)/\text{SE}$ is an asymptotic pivot for $l'\beta$. It depends upon the unknown parameters only through $l'\beta$, and it has a known limit distribution. This leads to a confidence interval for $l'\beta$.

The normal distribution is available in tables and in computer programs. We have

$$\text{Prob}(\mathcal{N}(0,1) > 1.96) = .025,$$

and since the normal distribution is symmetric about zero,

$$\text{Prob}(|\mathcal{N}(0,1)| \le 1.96) = .95.$$

Then Claim 3 gives

$$\lim_{n\to\infty} \text{Prob}(-1.96 \le \frac{l'\beta - l'b}{\text{SE}} \le 1.96) = .95,$$

and so

$$\lim_{n\to\infty} \text{Prob}(l'b - 1.96 \cdot \text{SE} \le l'\beta \le l'b + 1.96 \cdot \text{SE}) = .95,$$

or

$$\lim_{n\to\infty} \text{Prob}(\beta \in [l'b \pm 1.96 \cdot \text{SE}]) = .95.$$

4. HOMOSKEDASTIC CASE

So far, our basic assumption has simply been random sampling. In particular, we have *not* assumed that the conditional expectation is $X_i'\beta$ or that the conditional variance is constant. This is the power of asymptotics; we can do inference for a linear predictor when it is only an approximation to the conditional expectation function, and without restricting the form of the conditional variance function.

Now we are going to see how the asymptotic inference can be simplified if we do make these additional assumptions. So assume that

$$\text{(i) } E(Y_i \,|\, Z_i) = X_i'\beta \quad \text{and} \quad \text{(ii) } \text{Var}(Y_i \,|\, Z_i) = \sigma^2.$$

These assumptions are part of the normal linear model but we are not assuming that the conditional distribution of Y_i is normal. Stated in terms of the prediction error U_i, these assumptions are

$$\text{(i) } E(U_i \,|\, Z_i) = 0 \quad \text{and} \quad \text{(ii) } \text{Var}(U_i \,|\, Z_i) = E(U_i^2 \,|\, Z_i) = \sigma^2.$$

This implies that

$$\Sigma = E(U_1^2 X_1 X_1') = E[E(U_1^2 \mid Z_1) X_1 X_1'] = \sigma^2 E(X_1 X_1').$$

So there is a simpler form for the asymptotic covariance matrix:

$$\Lambda = [E(X_1 X_1')]^{-1} \sigma^2 E(X_1 X_1') [E(X_1 X_1')]^{-1} = \sigma^2 [E(X_1 X_1')]^{-1}.$$

Because $\sigma^2 = E(U_1^2)$, a consistent estimate of σ^2 is is SSR/n, and a consistent estimate of Λ is

$$\hat{\Lambda}^* = \frac{\mathrm{SSR}}{n} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1}.$$

The corresponding standard error is

$$\mathrm{SE}^* = (l' \hat{\Lambda}^* l / n)^{1/2} = [\frac{\mathrm{SSR}}{n} l' \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} l]^{1/2}.$$

Define the $n \times K$ matrix

$$X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix}.$$

Then

$$X'X = \sum_{i=1}^{n} X_i X_i',$$

and

$$\mathrm{SE}^* = [\frac{\mathrm{SSR}}{n} l' (X'X)^{-1} l]^{1/2}.$$

This coincides with the standard error used in Section 4 of Note 7 for our exact analysis of the normal linear model, except that the sum of squared residuals is divided by n instead of $n - K$. Usually this difference is not important, but it would be if K/n were bigger than, say, .2.