LECTURE NOTE 6

SAMPLING DISTRIBUTION

1. RANDOM SAMPLING

We have been able to discuss the population aspects of our models by considering a single draw from the population. That draw results in random variables, which have a joint distribution F:

$$(Y_1, \ldots, Y_M, Z_1, \ldots, Z_J) \sim F.$$

For example, with $M = 1$ and $J = 2$, we could have a population of people and for a randomly drawn individual, we have measures of Y = earnings, Z_1 = education, and Z_2 = labor market experience. With $M = 2$ and $J = 2$, we could have a population of families with twins, and for a randomly drawn family, Y_1 = earnings of twin 1, Y_2 = earnings of twin 2, Z_1 = education of twin 1, and Z_2 = education of twin 2. With $M = T$ and $J = 2T$, we could have a population of firms. For a randomly drawn firm, we have observations over T years, with Y_t = output in year t, Z_t = labor input in year t, and Z_{T+t} = capital input in year t ($t = 1, \ldots, T$).

We have discussed the use of data to form estimators of partial effects that have intuitive motivation, but we have not set up a formal link between the data and the population. This link is needed in order to obtain properties of our estimators and to construct confidence intervals that provide measures of uncertainty for the partial effects.

In a *random sample* of size n, we have n independent draws (with replacement) from the same population. The i^{th} draw results in the random variables

$$(Y_{i1}, \ldots, Y_{iM}, Z_{i1}, \ldots, Z_{iJ}).$$

The joint distribution of this list of random variables is the population distribution F. If $i \neq j$, the list of random variables for draw i is independent of the list for draw j, due to the random sampling. In general, of course, there is dependence within the list for i or within the list for j. We can summarize random sampling by saying

$$(Y_{i1}, \ldots, Y_{iM}, Z_{i1}, \ldots, Z_{iJ}) \overset{\text{i.i.d.}}{\sim} F \qquad (i = 1, \ldots, n).$$

Here "i.i.d." stands for "independent and identically distributed."

For notation, let

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iM} \end{pmatrix}, \quad Z_i = \begin{pmatrix} Z_{i1} \\ \vdots \\ Z_{iJ} \end{pmatrix}.$$

Then we can restate the random sampling condition as

$$(Y_i, Z_i) \overset{\text{i.i.d.}}{\sim} F \qquad (i = 1, \ldots, n).$$

In terms of the random sampling, there is no distinction between Y_i and Z_i. That distinction comes later, when we consider the conditional expectation of Y_i conditional on Z_i. For additional notation, define the $n \times M$ matrix of random variables

$$Y = \begin{pmatrix} Y_1' \\ \vdots \\ Y_n' \end{pmatrix} = \begin{pmatrix} Y_{11} & \ldots & Y_{1M} \\ \vdots & & \vdots \\ Y_{n1} & \ldots & Y_{nM} \end{pmatrix},$$

and the $n \times J$ matrix of random variables

$$Z = \begin{pmatrix} Z_1' \\ \vdots \\ Z_n' \end{pmatrix} = \begin{pmatrix} Z_{11} & \ldots & Z_{1J} \\ \vdots & & \vdots \\ Z_{n1} & \ldots & Z_{nJ} \end{pmatrix}.$$

Our linear predictors use X variables that are transformations of the Z variables:

$$X_{ik} = g_k(Z_i) \qquad (j = 1, \ldots, K),$$

using functions $g_k(\cdot)$ that we specify. For example, with $J = 1$, we could have $g_k(w) = w^{k-1}$, so that

$$X_{i1} = 1,\ X_{i2} = Z_i,\ \ldots,\ X_{iK} = Z_i^{K-1}.$$

The list of X variables will usually include a constant (for example, $X_{i1} = 1$), but our notation does not insist on including a constant. Let

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iK} \end{pmatrix}$$

and define the $n \times K$ random matrix

$$X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix} = \begin{pmatrix} X_{11} & \ldots & X_{1K} \\ \vdots & & \vdots \\ X_{n1} & \ldots & X_{nK} \end{pmatrix}.$$

Each element of this matrix is a function of Z, which we can write as $X = g(Z)$.

2. EXPECTATION OF THE LEAST-SQUARES ESTIMATOR

Consider the linear predictor

$$E^*(Y_i \mid X_i) = \beta_1 X_{i1} + \cdots + \beta_K X_{iK} = X_i'\beta$$

with $M = 1$, so that $Y_i = Y_{i1}$. The $K \times 1$ matrix β of coefficients is given by

$$\beta = [E(X_i X_i')]^{-1} E(X_i Y_i).$$

Note that β does not depend upon i, because the moments $E(X_i X_i')$ and $E(X_i Y_i)$ are the same for all i. This follows from the "identical" part of the i.i.d. condition for random sampling.

The least-squares estimator of β is

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

I am writing b as $b(Y, Z)$ to stress that it is a random variable, since it is a function $b(\cdot)$ evaluated at (Y, Z), and (Y, Z) is a random variable. (I shall often refer to matrices of random variables simply as random variables.)

I would like to calculate the expectation of b but this seems difficult because it is a nonlinear function (and, for example, $E[(X'X)^{-1}] \neq [E(X'X)]^{-1}$). The problem is simplified by working with the conditional distribution given $Z = z$. Then we can proceed as if $X = g(Z)$ were nonstochastic, with the value $x = g(z)$:

$$b(Y, Z) \,|\, Z = z \;\stackrel{d}{=}\; b(Y, z) \,|\, Z = z \;\stackrel{d}{=}\; (x'x)^{-1}x'Y \,|\, Z = z.$$

(Here $\stackrel{d}{=}$ means "has the same distribution as" or "equal in distribution.") This gives

$$E[b(Y, Z) \,|\, Z = z] = (x'x)^{-1}x'E(Y \,|\, Z = z).$$

The regression function $r(\cdot)$ is defined by

$$r(w) = E(Y_i \,|\, Z_i = w).$$

(This does not depend on i because the joint distribution of (Y_i, Z_i) is the same for all i.) Since (Y_i, Z_i) is independent of all the other elements in (Y, Z), we have

$$E(Y_i \,|\, Z = z) = E(Y_i \,|\, Z_i = z_i) = r(z_i),$$

and

$$E(b \,|\, Z = z) = (x'x)^{-1}x' \begin{pmatrix} r(z_1) \\ \vdots \\ r(z_n) \end{pmatrix}.$$

So far we have not made any assumptions other than random sampling. Now suppose that the linear predictor, which is intended to approximate the conditional expectation, actually equals the conditional expectation:

$$r(z_i) = x_i'\beta.$$

Then we have

$$E(b \mid Z = z) = (x'x)^{-1}x' \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix} \beta = (x'x)^{-1}(x'x)\beta = \beta.$$

Since this holds for any value of z, we can say

$$E(b \mid Z) = \beta,$$

and applying iterated expectations gives

$$E(b) = E[E(b \mid Z)] = E(\beta) = \beta$$

(provided that $E|b_k| < \infty$ for $k = 1, \dots, K$). So if $r(z_i) = x'_i\beta$, then b is an *unbiased* estimator for β.

If the linear predictor does not equal the conditional expectation, then b is not an unbiased estimator of the linear predictor coefficients. Nevertheless, $E(b \mid Z = z)$ does have a useful interpretation based on approximating the regression function. Consider the following approximation problem:

$$\gamma = \arg\min_d \sum_{i=1}^{n} [r(z_i) - x'_i d]^2.$$

So $x'_i\gamma$ provides an optimal approximation to $r(z_i)$, provided that we only consider the squared approximation errors at the z_i values that occur in our sample, and provided that we give these squared errors equal weight at all the z_i values. Then

$$\gamma = (x'x)^{-1}x' \begin{pmatrix} r(z_1) \\ \vdots \\ r(z_n) \end{pmatrix} = E(b \mid Z = z).$$

3. COVARIANCE MATRIX OF THE LEAST-SQUARES ESTIMATOR

We shall continue to work with the conditional distribution of Y given $Z = z$, so that $(x'x)^{-1}x'$ can be treated as nonrandom, with $x = g(z)$. Then applying Claim 1 from Note 5 gives

$$\mathrm{Cov}[b(Y, Z) \mid Z = z] = (x'x)^{-1}x'\mathrm{Cov}(Y \mid Z = z)x(x'x)^{-1}.$$

The covariance matrix of Y conditional on $Z = z$ is a $n \times n$ diagonal matrix:

$$\mathrm{Cov}(Y \mid Z = z) = \begin{pmatrix} \mathrm{Var}(Y_1 \mid Z_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathrm{Var}(Y_n \mid Z_n = z_n) \end{pmatrix}.$$

The off-diagonal terms are

$$\mathrm{Cov}(Y_i, Y_j \mid Z_i, Z_j) = 0,$$

because (Y_i, Z_i) and (Y_j, Z_j) are independent for $i \neq j$, due to random sampling. Define the *conditional variance function* $s(\cdot)$:

$$s(w) \equiv \mathrm{Var}(Y_i \mid Z_i = w).$$

Then we can write the conditional covariance matrix for Y given $Z = z$ as

$$\mathrm{Cov}(Y \mid Z = z) = \mathrm{diag}(s(z_1), \dots, s(z_n)).$$

So far we have not made any assumptions other than random sampling. Now introduce the assumption that the conditional variance function is constant:

$$s(w) = \mathrm{constant} \equiv \sigma^2,$$

with σ^2 defined as the value of this constant. The constant variance assumption is known as the *homoskedastic* case. When the conditional variance function is not constant, we say that there is *heteroskedasticity*. In the homoskedastic case, we have

$$\mathrm{Cov}(Y \mid \; = z) = \sigma^2 I_n,$$

where I_n is the $n \times n$ identity matrix. Then

$$\begin{aligned}
\mathrm{Cov}(b \,|\, Z = z) &= (x'x)^{-1} x' (\sigma^2 I_N) x (x'x)^{-1} \\
&= \sigma^2 (x'x)^{-1}.
\end{aligned}$$

$\mathrm{Cov}(b \,|\, Z = z)$ is $K \times K$:

$$\begin{aligned}
\mathrm{Cov}(b \,|\, Z = z) &= \begin{pmatrix} \mathrm{Cov}(b_1, b_1 \,|\, Z = z) & \dots & \mathrm{Cov}(b_1, b_K \,|\, Z = z) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(b_K, b_1 \,|\, Z = z) & \dots & \mathrm{Cov}(b_K, b_K \,|\, Z = z) \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{Var}(b_1 \,|\, Z = z) & \dots & \mathrm{Cov}(b_1, b_K \,|\, Z = z) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(b_K, b_1 \,|\, Z = z) & \dots & \mathrm{Var}(b_K \,|\, Z = z) \end{pmatrix}.
\end{aligned}$$

Let $[(x'x)^{-1}]_{jk}$ denote the (j,k) element of $(x'x)^{-1}$. Then we have

$$\mathrm{Cov}(b_j, b_k \,|\, Z = z) = \sigma^2 [(x'x)^{-1}]_{jk} \qquad (j, k = 1, \dots, K).$$

We can summarize our results as follows.

Claim 1. Under random sampling, if

$$E(Y_i \,|\, Z_i = z_i) = x_i' \beta \quad \text{and} \quad \mathrm{Var}(Y_i \,|\, Z_i = z_i) = \sigma^2,$$

then

$$E(b \,|\, Z = z) = \beta \quad \text{and} \quad \mathrm{Cov}(b \,|\, Z = z) = \sigma^2 (x'x)^{-1}.$$