Econometric Methods

Gary Chamberlain

This document contains the set of lecture notes from the late Gary Chamberlain's 2010 Econometrics class (EC2120) that I (Paul Goldsmith-Pinkham) took during my economics Ph.D. at Harvard University. Gary was a remarkable teacher and this class was an amazing experience for me as a young economist.

A few things worth noting from my experience taking this course:

- The course is somewhat unique in not introducing any inference until Lecture 7 (halfway through the course). The focus prior to this is exclusively on estimation using regression.
- The lectures are linked together in groups (even though they are not marked this way). Lectures 1-3 reflect the underlying setup in notation and framework for the rest of the course. Lectures 6-9 setup inference. Lecture 9-11 discuss a general framework for using moment conditions.
- During my semester, we never got through lectures 13-15, which suggests that this is a lot of material for a single semester.
- Gary would continually refer back to Lecture 4 and Lecture 9 (indeed, I have Gary's voice saying "if we think back to the Note 9 framework..." burned into my brain. As a result, these are also some of the longest and densest lectures.
- The last 4 sections are review problems that Gary provided for preparation for the final exam. These are very fun problems, but we were not provided solutions. So, you'll have to figure them out on your own!

- Paul Goldsmith-Pinkham (Harvard Ph.D., 2015)

Contents

LECTURE NOTE 1

LINEAR PREDICTOR AND LEAST-SQUARES FIT

1. LINEAR PREDICTOR

Consider a random sample of n individuals that provides data on their earnings and education. Consider the first individual in the sample, and let Y denote her earnings and let X denote her education. I want you to think of (Y, X) as a pair of *random variables*. The randomness comes from the act of random sampling: before this individual is drawn from the population, we do not know what the earnings and education will turn out to be, but we can assign a joint distribution to (Y, X).

It would be nice if there were a function connecting Y and X: $Y = f(X)$, but no, individuals with the same education may have different earnings. A more promising goal is to establish a relationship in a predictive sense. Given the value of X, we can try to predict the value of Y, and a good place to start is a *linear predictor*:

$$\hat{Y} = \beta_0 + \beta_1 X.$$

Now we have to say how those coefficients β_0 and β_1 are going to get determined. A very convenient criterion is the square of the prediction error, and we choose β_0 and β_1 to minimize its expectation:

$$\min_{\beta_0, \beta_1} E(Y - \hat{Y})^2.$$

So a more complete description of our linear predictor is *minimum mean square error* linear predictor.

Similar minimization problems come up elsewhere in the course, and on the principle that "the same equations have the same solutions," I'd like to once and for all lay out a

way to solve these problems. The key is to use *orthogonal projection* in a vector space with an *inner product*. Here the inner product is

$$\langle Y, X \rangle = E(YX).$$

The associated *norm* is

$$||Y|| = \langle Y, Y \rangle^{1/2}.$$

Then we can restate our linear predictor problem as

$$\min_{\beta_0, \beta_1} ||Y - \hat{Y}||^2.$$

The solution is obtained from the orthogonal projection of Y on 1 and X. It is convenient to define X_0 as a degenerate random variable that only takes on the value 1. The orthogonal projection requires that the prediction error $(Y - \hat{Y})$ is orthogonal to X_0 and X:

$$\langle Y - \hat{Y}, X_0 \rangle = 0,$$
$$\langle Y - \hat{Y}, X \rangle = 0.$$

Notation for this orthogonality is

$$Y - \hat{Y} \perp X_0, \quad Y - \hat{Y} \perp X.$$

Writing out the two orthogonality conditions gives

$$\langle Y - \beta_0 X_0 - \beta_1 X, X_0 \rangle = \langle Y, X_0 \rangle - \beta_0 \langle X_0, X_0 \rangle - \beta_1 \langle X, X_0 \rangle = 0,$$
$$\langle Y - \beta_0 X_0 - \beta_1 X, X \rangle = \langle Y, X \rangle - \beta_0 \langle X_0, X \rangle - \beta_1 \langle X, X \rangle = 0.$$

Using our definition for the inner product,

$$E(Y) - \beta_0 - \beta_1 E(X) = 0,$$
$$E(YX) - \beta_0 E(X) - \beta_1 E(X^2) = 0.$$

This gives two linear equations for the two unknowns, β_0 and β_1. These equations can be solved to give

$$\beta_1 = \frac{E(YX) - E(Y)E(X)}{E(X^2) - E(X)E(X)}$$
$$\beta_0 = E(Y) - \beta_1 E(X).$$

The numerator in the expression for β_1 can be taken as the definition of *covariance*:

$$\mathrm{Cov}(Y, X) \equiv E(YX) - E(Y)E(X),$$

and the denominator can be taken as the definition of *variance*:

$$\mathrm{Var}(X) \equiv E(X^2) - E(X)E(X).$$

So we can rewrite the slope coefficient in the linear predictor as

$$\beta_1 = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)}.$$

Our notation for the (population) linear predictor is

$$E^*(Y \mid 1, X) = \beta_0 + \beta_1 X.$$

2. LEAST-SQUARES FIT

The data from a sample of size n can be put into two matrices:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

and it convenient to define an additional matrix x_0, which is simply a $n \times 1$ column of 1's:

$$x_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The fitted value for the i^{th} observation is

$$\hat{y}_i = b_0 + b_1 x_i,$$

and the objective is to choose the coefficients b_0 and b_1 to minimize the sum of squared residuals:

$$\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

(Dividing by n is not necessary, but it does suggest an analogy with minimizing mean square error in the population.)

Define the inner product

$$\langle y, x \rangle = \frac{1}{n} \sum_{i=1}^{n} y_i x_i.$$

Now we have a minimum norm problem:

$$\min_{b_0, b_1} ||y - b_0 x_0 - b_1 x||^2,$$

and the solution, once again, is obtained from the orthogonal projection of y on x_0 and x. This requires that the prediction error $(y - \hat{y})$ be orthogonal to x_0 and x:

$$\langle y - \hat{y}, x_0 \rangle = 0,$$

$$\langle y - \hat{y}, x \rangle = 0.$$

Notation for this orthogonality is

$$y - \hat{y} \perp x_0, \quad y - \hat{y} \perp x.$$

Writing out the orthogonality conditions gives

$$\langle y, x_0 \rangle - b_0 \langle x_0, x_0 \rangle - b_1 \langle x, x_0 \rangle = 0,$$

$$\langle y, x \rangle - b_0 \langle x_0, x \rangle - b_1 \langle x, x \rangle = 0.$$

Using our definition for the (least-squares) inner product, we have

$$\bar{y} - b_0 - b_1\bar{x} = 0,$$

$$\overline{yx} - b_0\bar{x} - b_1\overline{x^2} = 0,$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{yx} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

The two linear equations for the two unknowns, b_0 and b_1, can be solved to give

$$b_1 = \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}\bar{x}},$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

Our notation for the least-squares fit (or sample linear predictor) is

$$\hat{y}_i \mid 1, x = b_0 + b_1 x_i.$$

3. GOODNESS OF FIT

Note that

$$0 \leq \frac{||Y - E^*(Y \mid 1, X)||^2}{||Y - E^*(Y \mid 1)||^2} \leq 1.$$

This ratio is less than or equal to 1 because using X to predict Y cannot increase the mean square error—β_1 is allowed to be 0. (The linear predictor using just a constant is $E^*(Y \mid 1) = E(Y)$.) We define a measure of goodness of fit in the population as

$$R^2_{\text{pop}} = 1 - \frac{||Y - E^*(Y \mid 1, X)||^2}{||Y - E^*(Y \mid 1)||^2}.$$

This measure is scale free in that it is not affected if Y is multiplied by a constant (for example, changing the units from dollars to cents). It is easy to interpret since

$$0 \leq R^2_{\text{pop}} \leq 1.$$

The sample counterpart is

$$R^2 = 1 - \frac{||y - (\hat{y} \,|\, 1, x)||^2}{||y - (\hat{y} \,|\, 1)||^2}.$$

(The least-squares fit using just a constant is $(\hat{y} \,|\, 1) = \bar{y}$.) It is also scale free with

$$0 \leq R^2 \leq 1.$$

4. OMITTED VARIABLES

Consider an individual chosen at random from a population. Let Y denote her earnings, and let X_1 and X_2 denote her education and her score on a test administered when she was in the third grade. The random variables (Y, X_1, X_2) have a joint distribution. There is a (population) linear predictor for Y given X_1 and X_2 (and a constant):

$$E^*(Y \,|\, 1, X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \qquad (Long)$$

and there is a (population) linear predictor for Y just given X_1 (and a constant):

$$E^*(Y \,|\, 1, X_1) = \alpha_0 + \alpha_1 X_1. \qquad (Short)$$

I want to develop the relationship between these two linear predictors. This requires the auxiliary linear predictor of X_2 given X_1 (and a constant):

$$E^*(X_2 \,|\, 1, X_1) = \gamma_0 + \gamma_1 X_1. \qquad (Aux)$$

Let U denote the prediction error using the long predictor:

$$U \equiv Y - E^*(Y \,|\, 1, X_1, X_2),$$

so that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U. \qquad (1)$$

Because U is a prediction error, it is orthogonal to the variables used in the predictor:

$$U \perp 1, \quad U \perp X_1, \quad U \perp X_2.$$

In particular, U is orthogonal to 1, X_1, which implies that

$$E^*(U \mid 1, X_1) = 0. \tag{2}$$

Use equations (1) and (2) to write the short predictor as

$$\begin{aligned} E^*(Y \mid 1, X_1) &= \beta_0 + \beta_1 X_1 + \beta_2 E^*(X_2 \mid 1, X_1) + E^*(U \mid 1, X_1) \\ &= \beta_0 + \beta_1 X_1 + \beta_2(\gamma_0 + \gamma_1 X_1) + 0 \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)X_1. \end{aligned}$$

So we have proved the following

Claim 1. $\alpha_0 = \beta_0 + \beta_2\gamma_0, \quad \alpha_1 = \beta_1 + \beta_2\gamma_1.$

The coefficient α_1 on X_1 in the short predictor is the coefficient β_1 from the long predictor plus an additional term. This additional term is the product of the coefficient β_2 on the omitted variable and the coefficient γ_1 on X_1 in the auxiliary predictor. This result is often called the *omitted variable bias* formula. If the goal is the coefficient on X_1 in the linear predictor that includes X_1 and X_2, then the coefficient on X_1 in the short predictor differs from this goal by $\beta_2\gamma_1$. Note that this bias term is 0 if $\gamma_1 = 0$, which holds if $\mathrm{Cov}(X_1, X_2) = 0$.

There is a similar result for the least-squares fit using sample data. Our notation for the long, short, and auxiliary least-squares fit is

$$\begin{aligned} \hat{y}_i \mid 1, x_{i1}, x_{i2} &= b_0 + b_1 x_{i1} + b_2 x_{i2}, \\ \hat{y}_i \mid 1, x_{i1} &= a_0 + a_1 x_{i1}, \\ \hat{x}_{i2} \mid 1, x_{i1} &= c_0 + c_1 x_{i1}. \end{aligned}$$

The argument above using the population predictors translates directly into an argument using sample predictors (least-squares fits). Just change the inner product from $E(XY)$ to $\sum_{i=1}^{n} y_i x_i / n$. This gives

Claim 2. $a_0 = b_0 + b_2 c_0, \quad a_1 = b_1 + b_2 c_1$.

This least-squares version of the omitted variable bias formula is a computational identity, which can be checked on a data set using a least-squares computer program.

LECTURE NOTE 2

CONDITIONAL EXPECTATION

1. FUNCTIONAL FORM

The linear predictor is very flexible because we are free to construct transformations of the original variables. For example, if EXP is a measure of years of job market experience, we can set $X_1 = \text{EXP}$ and $X_2 = \text{EXP}^2$. Then evaluating

$$E^*(Y \mid 1, X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

at $\text{EXP} = c$ gives

$$\beta_0 + \beta_1 c + \beta_2 c^2,$$

and we can do this evaluation for several interesting values for experience.

The same point applies with two or more original variables. Suppose that in addition to EXP we have EDUC, a measure of years of education. We can set $X_1 = \text{EDUC}$, $X_2 = \text{EXP}$, and $X_3 = \text{EDUC} \cdot \text{EXP}$. Then evaluating

$$E^*(Y \mid 1, X_1, X_2, X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

at $\text{EDUC} = c$ and $\text{EXP} = d$ gives

$$\beta_0 + \beta_1 c + \beta_2 d + \beta_3 c \cdot d,$$

and we can do this evaluation for several interesting values for education and experience.

2. CONDITIONAL EXPECTATION

Suppose that we start with a single original variable Z and develop linear predictors of Y based on Z that are increasingly flexible. To be specific, consider using a polynomial of order M:

$$E^*(Y \mid 1, Z, Z^2, \ldots, Z^M).$$

The expectation of the squared prediction error cannot increase as M increases, because the coefficients on the additional terms are allowed to be 0. So

$$E[Y - E^*(Y \mid 1, Z, Z^2, \ldots, Z^M)]^2$$

is decreasing as $M \to \infty$ and must approach a limit (since it is nonnegative). We shall assume that the linear predictor itself approaches a limit, and we shall identify this limit with the conditional expectation, $E(Y \mid Z)$:

$$E(Y \mid Z) = \lim_{M \to \infty} E^*(Y \mid 1, Z, Z^2, \ldots, Z^M).$$

This limit is in a mean square sense:

$$\lim_{M \to \infty} E[E(Y \mid Z) - E^*(Y \mid 1, Z, Z^2, \ldots, Z^M)]^2 = 0.$$

We can think of the conditional expectation as providing the best prediction of Y given Z, with (essentially) no constraint on the functional form of the predictor.

Let U be notation for the prediction error:

$$U \equiv Y - E(Y \mid Z).$$

Then U is orthogonal to any power of Z:

$$\langle U, Z^j \rangle = E(UZ^j) = 0 \qquad (j = 0, 1, 2, \ldots).$$

Because general functions of Z can be approximated (in mean square) by polynomials in Z, we have

$$\langle U, g(Z) \rangle = E[Ug(Z)] = 0$$

for (essentially) arbitrary functions $g(\cdot)$.

In the population, we shall generally prefer to work with the conditional expectation. The linear predictor remains useful, however, because it has a direct sample counterpart: the sample linear predictor or least-squares fit. We shall use a (population) linear predictor to approximate the conditional expectation, and then use use a least squares fit to estimate the linear predictor.

It is useful to have notation for evaluating the conditional expectation at a particular value for Z:

$$r(z) \equiv E(Y \mid Z = z).$$

The function $r(\cdot)$ is called the *regression function*. The regression function evaluated at the random variable Z is the conditional expectation: $r(Z) = E(Y \mid Z)$. Because the regression function may be complicated, we may want to approximate it by a simpler function that would be easier to estimate. For example, $E^*[r(Z) \mid 1, Z]$ is a minimum mean-square error approximation that uses a linear function of Z. This turns out to be the same as the linear predictor of Y given Z:

Claim 1. $E^*[r(Z) \mid 1, Z] = E^*(Y \mid 1, Z) = \beta_0 + \beta_1 Z$.

Proof. Let U denote the prediction error:

$$U \equiv Y - E(Y \mid Z) = Y - r(Z). \tag{1}$$

Then U is orthogonal to any function of Z:

$$E[Ug(Z)] = 0,$$

and so is orthogonal to 1 and to Z:

$$E(U) = E(UZ) = 0.$$

This implies that the linear predictor of U given 1, Z is 0, and applying that to (1) gives

$$0 = E^*(U \mid 1, Z) = E^*(Y \mid 1, Z) - E^*[r(Z) \mid 1, Z]. \quad \diamond$$

The conditional expectation of Y given two (or more) variables Z_1 and Z_2 can also be viewed as a limit of increasingly flexible linear predictors:

$$E(Y \mid Z_1, Z_2) = \lim_{M \to \infty} E^*(Y \mid 1, Z_1, Z_2, Z_1^2, Z_1 Z_2, Z_2^2, \ldots, Z_1^M, Z_1^{M-1} Z_2, \ldots, Z_1 Z_2^{M-1}, Z_2^M).$$

The regression function is defined as

$$r(z_1, z_2) \equiv E(Y \mid Z_1 = z_1, Z_2 = z_2).$$

As above, we can use the linear predictor to approximate the regression function. For example, the proof of claim 1 can be used to show that

$$E^*[r(Z_1, Z_2) \mid 1, Z_1, Z_2, Z_1^2, Z_1 Z_2, Z_2^2] = E^*(Y \mid 1, Z_1, Z_2, Z_1^2, Z_1 Z_2, Z_2^2).$$

We shall conclude this section by deriving the iterated expectations formula and then using it to obtain an omitted variables formula.

Claim 2 (Iterated Expectations). $E[E(Y \mid Z_1, Z_2) \mid Z_1] = E(Y \mid Z_1)$.
(Equivalently: $E[r(Z_1, Z_2) \mid Z_1] = r(Z_1)$.)
Proof. Let U denote the prediction error:

$$U \equiv Y - E(Y \mid Z_1, Z_2) = Y - r(Z_1, Z_2). \tag{2}$$

Then U is orthogonal to any function of (Z_1, Z_2):

$$E[U g(Z_1, Z_2)] = 0,$$

and so is orthogonal to any function of Z_1:

$$E[U g(Z_1)] = 0.$$

This implies that $E(U \mid Z_1) = 0$, and applying that to (2) gives

$$0 = E(U \mid Z_1) = E(Y \mid Z_1) - E[r(Z_1, Z_2) \mid Z_1]. \quad \diamond$$

Claim 3 (Omitted Variable Bias). If

$$E(Y \mid Z_1, Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$$

and

$$E(Z_2 \mid Z_1) = \gamma_0 + \gamma_1 Z_1,$$

then

$$E(Y \mid Z_1) = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)Z_1.$$

Proof.

$$\begin{aligned} E(Y \mid Z_1) &= E[E(Y \mid Z_1, Z_2) \mid Z_1] \\ &= E(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 \mid Z_1) \\ &= \beta_0 + \beta_1 Z_1 + \beta_2 E(Z_2 \mid Z_1) \\ &= \beta_0 + \beta_1 Z_1 + \beta_2(\gamma_0 + \gamma_1 Z_1) \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)Z_1. \quad \diamond \end{aligned}$$

Note that here we assume that the regression function for Y on Z_1 and Z_2 is linear in Z_1 and Z_2, and that the regression function for Z_2 on Z_1 is linear in Z_1. It then follows that the regression function for Y on Z_1 is linear in Z_1, and the coefficients are related to the coefficients in the long regression function in the same way as in claim 1 in Note 1.

3. DISCRETE REGRESSORS

Suppose that Z_1 and Z_2 take on only a finite set of values:

$$Z_1 \in \{\lambda_1, \ldots, \lambda_J\}, \quad Z_2 \in \{\delta_1, \ldots, \delta_K\}.$$

Construct the following *dummy variables*:

$$\begin{aligned} X_{jk} &= \begin{cases} 1, & \text{if } Z_1 = \lambda_j,\ Z_2 = \delta_k; \\ 0, & \text{otherwise;} \end{cases} \\ &= 1(Z_1 = \lambda_j, Z_2 = \delta_k) \qquad (j = 1, \ldots, J;\ k = 1, \ldots, K). \end{aligned}$$

These are indicator variables that equal 1 if a particular value of (Z_1, Z_2) occurs, and equal 0 otherwise. We use the notation $1(B)$ for the indicator function that equals 1 if the event B occurs and equals 0 otherwise.

Claim 4. $E(Y \mid Z_1, Z_2) = E^*(Y \mid X_{11}, \ldots, X_{J1}, \ldots, X_{1K}, \ldots, X_{JK})$

Proof. Any function $g(Z_1, Z_2)$ can be written as

$$g(Z_1, Z_2) = \sum_{j=1}^{J} \sum_{k=1}^{K} \gamma_{jk} X_{jk}$$

with $\gamma_{jk} = g(\lambda_j, \delta_k)$. So searching over functions g to find the best predictor is equivalent to searching over the coefficients γ_{jk} to find the best linear predictor. $\diamond$

So the conditional expectation function can be expressed as a linear combination of the dummy variables:

$$E(Y \mid Z_1, Z_2) = \sum_{j=1}^{J} \sum_{k=1}^{K} \beta_{jk} X_{jk}$$

with

$$\beta_{jk} = E(Y \mid Z_1 = \lambda_j, Z_2 = \delta_k).$$

Note this requires that we use a complete set of dummy variables, with one for each value of (Z_1, Z_2). In this discrete regressor case, there is a concrete form for the notion that conditional expectation is a limit of increasingly flexible linear predictors. Here the limit is achieved by using a complete set of dummy variables in the linear predictor.

There is a sample analog to this result, using least-squares fits. The basic data consist of (y_i, z_{i1}, z_{i2}) for each of $i = 1, \ldots, n$ members of the sample. Construct the dummy variables

$$x_{i,jk} = 1(z_{i1} = \lambda_j, z_{i2} = \delta_k)$$

and the matrices

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad x_{jk} = \begin{pmatrix} x_{1,jk} \\ \vdots \\ x_{n,jk} \end{pmatrix} \qquad (j = 1, \ldots, J;\ k = 1, \ldots, K).$$

The coefficients in the least-squares fit are obtained from

$$\min ||y - \sum_{j=1}^{J} \sum_{k=1}^{K} b_{jk} x_{jk}||^2,$$

where the minimization is over $\{b_{jk}\}$ and the inner product is

$$\langle y, x_{jk} \rangle = \frac{1}{n} \sum_{i=1}^{n} y_i x_{i,jk}.$$

Claim 5.

$$b_{lm} = \frac{\sum_{i=1}^{n} y_i x_{i,lm}}{\sum_{i=1}^{n} x_{i,lm}} \qquad (l = 1, \ldots, J;\ m = 1, \ldots, K).$$

Proof. The residual from the least-squares fit must be orthogonal to each of the dummy variables:

$$\langle y - \sum_{j,k} b_{jk} x_{jk}, x_{lm} \rangle = 0.$$

The dummy variables are orthogonal to each other:

$$\langle x_{jk}, x_{lm} \rangle = 0$$

unless $j = l$ and $k = m$. So we have

$$\begin{aligned} \langle y - \sum_{j,k} b_{jk} x_{jk}, x_{lm} \rangle &= \langle y, x_{lm} \rangle - \sum_{j,k} b_{j,k} \langle x_{jk}, x_{lm} \rangle \\ &= \langle y, x_{lm} \rangle - b_{lm} \langle x_{lm}, x_{lm} \rangle \\ &= 0. \end{aligned}$$

So

$$b_{lm} = \frac{\langle y, x_{lm} \rangle}{\langle x_{lm}, x_{lm} \rangle} = \frac{\sum_{i=1}^{n} y_i x_{i,lm}}{\sum_{i=1}^{n} x_{i,lm}}.$$

(Note that $x_{i,lm}^2 = x_{i,lm}$ because $x_{i,lm}$ equals 0 or 1.) $\diamond$

Note that $\sum_i y_i x_{i,lm}$ is summing the y values for the observations with $(z_{i1}, z_{i2}) = (\lambda_l, \delta_m)$, and $\sum_i x_{i,lm}$ is the number of observations with this value for (z_{i1}, z_{i2}). So the

coefficient b_{lm} is a subsample mean, for the subsample with $(z_{i1}, z_{i2}) = (\lambda_l, \delta_m)$. In order to stress this interpretation as a subsample mean, we shall use the notation

$$\bar{y} \,|\, \lambda_l, \delta_m \equiv \frac{\sum_{i=1}^n y_i x_{i,lm}}{\sum_{i=1}^n x_{i,lm}}.$$

A major use of regression analysis is to measure the effect of one variable holding constant other variables. Consider, for example, the effect on Y of a change from $Z_1 = c$ to $Z_1 = d$, holding Z_2 constant at $Z_2 = e$. Let θ denote this effect:

$$\begin{aligned}\theta &= E(Y \,|\, Z_1 = d, Z_2 = e) - E(Y \,|\, Z_1 = c, Z_2 = e) \\ &= r(d, e) - r(c, e).\end{aligned}$$

This is a predictive effect. It measures how the prediction of Y changes as we change the value for one of the predictor variables, holding constant the value of the other predictor variable. In the case of discrete regressors with a complete set of dummy variables, this predictive effect has a sample analog:

$$\hat{\theta} = (\bar{y} \,|\, d, e) - (\bar{y} \,|\, c, e).$$

We estimate θ by comparing two subsample means. The individuals in the first subsample have $z_{i1} = c$, and the individuals in the second subsample have $z_{i1} = d$. In both subsamples, all individuals have the same value for z_2: $z_{i2} = e$. So the sense in which z_2 is being held constant is clear: all individuals in the comparison of means have the same value for z_2.

In general there is a different effect θ for each value of Z_2, and we may want to have a way to summarize these effects. This is discussed in the next section.

4. AVERAGE PARTIAL EFFECT

Recall our definition of the regression function:

$$r(s, t) = E(Y \,|\, Z_1 = s, Z_2 = t).$$

Consider the predictive effect based on comparing $Z_1 = c$ with $Z_1 = d$, with $Z_2 = t$:

$$r(d,t) - r(c,t).$$

Instead of reporting a different effect for each value of Z_2, we can evaluate the effect at the random variable Z_2:

$$r(d,Z_2) - r(c,Z_2).$$

This gives a random variable, and we can take its expectation:

$$\theta = E[r(d,Z_2) - r(c,Z_2)].$$

We shall refer to this as an *average partial effect*. It is "partial" in the sense of holding Z_2 constant.

Note that

$$\theta = E[r(d,Z_2)] - E[r(c,Z_2)].$$

This is not the same, in general, as

$$\begin{aligned}\tilde{\theta} &= E[r(d,Z_2) \mid Z_1 = d] - E[r(c,Z_2) \mid Z_1 = c] \\ &= E(Y \mid Z_1 = d) - E(Y \mid Z_1 = c).\end{aligned}$$

An integral notation may be helpful. Define

$$\begin{aligned}F(B \mid s,t) &= \text{Prob}(Y \in B \mid Z_1 = s, Z_2 = t), \\ F_{Z_2}(B) &= \text{Prob}(Z_2 \in B), \\ F_{Z_2}(B \mid s) &= \text{Prob}(Z_2 \in B \mid Z_1 = s).\end{aligned}$$

Then

$$r(s,t) = \int y dF(y \mid s,t),$$

and

$$\theta = \int [r(d,t) - r(c,t)] dF_{Z_2}(t).$$

We have $\theta = \tilde{\theta}$ if

$$F_{Z_2}(B \mid c) = F_{Z_2}(B \mid d) = F_{Z_2}(B)$$

for all B; $\theta = \tilde{\theta}$ for all c and d if Z_1 and Z_2 are independent.

Once we have an estimate $\hat{r}$ of the regression function, we can form an estimate of θ by taking an average over the sample:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} [\hat{r}(d, z_{i2}) - \hat{r}(c, z_{i2})]. \tag{3}$$

We can obtain an estimate of r by first approximating the conditional expectation by a linear predictor, using a polynomial in Z_1 and Z_2:

$$\begin{aligned} E(Y \mid Z_1, Z_2) &\cong E^*(Y \mid \{Z_1^j \cdot Z_2^k\}_{j+k=0}^{M}) \\ &= \sum_{j,k:j+k=0}^{M} \beta_{jk} Z_1^j \cdot Z_2^k. \end{aligned}$$

We can use a least-squares fit to obtain estimates b_{jk} of the coefficients β_{jk}. Then we can use

$$\hat{r}(c, z_{i2}) = \sum_{j,k:j+k=0}^{M} b_{jk} c^j \cdot z_{i2}^k \quad \text{and} \quad \hat{r}(d, z_{i2}) = \sum_{j,k:j+k=0}^{M} b_{jk} d^j \cdot z_{i2}^k$$

in (3).

Now consider the special case of discrete regressors, with

$$Z_1 \in \{\lambda_1, \ldots, \lambda_J\}, \quad Z_2 \in \{\delta_1, \ldots, \delta_K\}.$$

In this case,

$$\theta = \sum_{k=1}^{K} [r(d, \delta_k) - r(c, \delta_k)] \text{Prob}(Z_2 = \delta_k).$$

We can estimate θ using the sample analog

$$\hat{\theta} = \sum_{k=1}^{K} [(\bar{y} \mid d, \delta_k) - (\bar{y} \mid c, \delta_k)](n_k/n),$$

where n_k is the number of observations with $z_{i2} = \delta_k$:

$$n_k = \sum_{i=1}^{n} 1(z_{i2} = \delta_k),$$

and the mean of y for a subsample is

$$(\bar{y} \mid s, t) \equiv \frac{\sum_{i=1}^{n} y_i 1(z_{i1} = s, z_{i2} = t)}{\sum_{i=1}^{n} 1(z_{i1} = s, z_{i2} = t)}.$$

5. LOGS

Section 1 stressed that the linear predictor is flexible because we are free to construct transformations of the original variables. A transformation that is often used is the logarithm:

$$E^*(Y \mid 1, \log Z) = \beta_0 + \beta_1 \log Z.$$

(This is the log to the base e or natural logarithm ln.) In order to compare $Z = c$ and $Z = d$, we simply substitute:

$$\beta_1 \log d - \beta_1 \log c = \beta_1 \log(d/c).$$

A useful approximation here is

$$(\beta_1/100)[100 \log(d/c)] \cong (\beta_1/100)[100(\frac{d}{c} - 1)].$$

With this approximation, we can interpret $(\beta_1/100)$ as the (predictive) effect of a one per cent change in Z.

Now consider a log transformation of Y:

$$E^*(\log Y \mid 1, Z) = \beta_0 + \beta_1 Z.$$

We can certainly say that the predicted change in $\log Y$ is $\beta_1(d - c)$, and it is often useful to think of $100\beta_1(d - c)$ as a predicted percentage change in Y. We should note, however, that even if the conditional expectation of $\log Y$ is linear, so that

$$E(\log Y \mid Z) = \beta_0 + \beta_1 Z,$$

we cannot relate this to the conditional expectation of Y without additional assumptions.

To see this, define

$$U \equiv \log Y - E(\log Y \mid Z),$$

so that

$$E(U \mid Z) = 0.$$

Since $\log Y = \beta_0 + \beta_1 Z + U$, we have

$$\begin{aligned} Y &= \exp(\beta_0 + \beta_1 Z + U) \\ &= \exp(\beta_0 + \beta_1 Z) \cdot \exp(U). \end{aligned}$$

So

$$E(Y \mid Z) = \exp(\beta_0 + \beta_1 Z) \cdot E[\exp(U) \mid Z].$$

In general, $E(U \mid Z) = 0$ does not imply that $E[\exp(U) \mid Z]$ is a constant. If we make an additional assumption that U and Z are independent, then

$$E[\exp(U) \mid Z] = E[\exp(U)].$$

In that case,

$$\frac{E(Y \mid Z = d)}{E(Y \mid Z = c)} = \exp[\beta_1(d - c)] \cong \beta_1(d - c) + 1,$$

and

$$100\left[\frac{E(Y \mid Z = d)}{E(Y \mid Z = c)} - 1\right] \cong 100\beta_1(d - c).$$

LECTURE NOTE 3

RESIDUAL REGRESSION, OMITTED VARIABLES, AND A MATRIX VERSION

1. RESIDUAL REGRESSION

Consider the linear predictor with a general list of K predictor variables (plus a constant):

$$E^*(Y \mid 1, X_1, \ldots, X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K. \tag{1}$$

We are going to develop a formula for a single coefficient, which, for convenience, will be β_K. Our result will use the linear predictor of X_K given the other predictor variables:

$$E^*(X_K \mid 1, X_1, \ldots, X_{K-1}) = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{K-1} X_{K-1}.$$

Define $\tilde{X}_K$ as the residual (prediction error) from this linear predictor:

$$\tilde{X}_K = X_K - E^*(X_K \mid 1, X_1, \ldots, X_{K-1}).$$

The result is that β_K is the coefficient on $\tilde{X}_K$ in the linear predictor of Y given just $\tilde{X}_K$:

Claim 1. $E^*(Y \mid \tilde{X}_K) = \beta_K \tilde{X}_K$ with $\beta_K = E(Y\tilde{X}_K)/E(\tilde{X}_K^2)$.

Proof. Substitute

$$X_K = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{K-1} X_{K-1} + \tilde{X}_K$$

into (1) to obtain

$$\begin{aligned} E^*(Y \mid 1, X_1, \ldots, X_K) &= \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} \\ &\quad + \beta_K(\gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{K-1} X_{K-1} + \tilde{X}_K) \\ &= \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_{K-1} X_{K-1} + \beta_K \tilde{X}_K, \end{aligned} \tag{2}$$

with

$$\tilde{\beta}_j = \beta_j + \beta_K \gamma_j \qquad (j = 0, 1, \ldots, K-1). \tag{3}$$

The residual from predicting Y must be orthogonal to $1, X_1, \ldots, X_K$. Since $\tilde{X}_K$ is a linear combination of $1, X_1, \ldots, X_K$, we must have $\tilde{X}_k$ orthogonal to $Y - E^*(Y \mid 1, X_1, \ldots, X_K)$:

$$\langle Y - \tilde{\beta}_0 - \tilde{\beta}_1 X_1 - \cdots - \tilde{\beta}_{K-1} X_{K-1} - \beta_K \tilde{X}_K, \tilde{X}_K \rangle = 0. \tag{4}$$

Since $\tilde{X}_K$ is the residual from a prediction based on $1, X_1 \ldots, X_{K-1}$, it is orthogonal to those variables, and (4) reduces to

$$\langle Y - \beta_K \tilde{X}_K, \tilde{X}_K \rangle = \langle Y, \tilde{X}_K \rangle - \beta_K \langle \tilde{X}_K, \tilde{X}_K \rangle = 0.$$

So $\beta_K \tilde{X}_K$ is the orthogonal projection of Y on $\tilde{X}_K$ and

$$\beta_K = \langle Y, \tilde{X}_K \rangle / \langle \tilde{X}_K, \tilde{X}_K \rangle = E(Y \tilde{X}_K) / E(\tilde{X}_K^2). \quad \diamond$$

This population result has a sample counterpart. The data are in the matrices

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad x_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \qquad (j = 0, 1, \ldots, K).$$

Consider the least-squares fit using the K predictor variables (and a constant):

$$\hat{y}_i \mid 1, x_1, \ldots, x_K = b_0 + b_1 x_{i1} + \ldots + b_K x_{iK}.$$

Claim 2 provides a formula for a single coefficient, such as b_K. The result uses the least-squares fit of x_K on the other predictor variables:

$$\hat{x}_{iK} \mid 1, x_1, \ldots, x_{K-1} = c_0 + c_1 x_{i1} + \cdots + c_{K-1} x_{i,K-1}.$$

Define $\tilde{x}_K$ as the residual from this least-squares fit:

$$\tilde{x}_{iK} = x_{iK} - (\hat{x}_{iK} \mid 1, x_1, \ldots, x_{K-1}).$$

Then b_K is the coefficient on $\tilde{x}_K$ in the least squares fit of y on just $\tilde{x}_K$:

Claim 2. $(\hat{y}_i \,|\, \tilde{x}_K) = b_k \tilde{x}_{iK}$ with $b_K = \frac{1}{n}\sum_{i=1}^n y_i \tilde{x}_{iK} \Big/ \frac{1}{n}\sum_{i=1}^n \tilde{x}_{iK}^2$.

The proof is the same as for claim 1, with the least-squares inner product $\langle y, x_j \rangle = \sum_{i=1}^n y_i x_{ij}/n$ replacing the linear predictor (or mean-square) inner product $\langle Y, X_j \rangle = E(Y X_j)$.

2. OMITTED VARIABLES

This section derives the general version, with K predictor variables, of the omitted variable formula in claim 1 of Note 1. We shall use the notation (and part of the argument) from the residual regression result in Section 1. The short linear predictor is

$$E^*(Y \,|\, 1, X_1, \ldots, X_{K-1}) = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{K-1} X_{K-1}.$$

Claim 3. $\alpha_j = \beta_j + \beta_K \gamma_j \qquad (j = 0, 1, \ldots, K-1).$

Proof. Let U denote the following prediction error:

$$U \equiv Y - E^*(Y \,|\, 1, X_1, \ldots, X_K).$$

Use equation (2) to write

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_{K-1} X_{K-1} + \beta_K \tilde{X}_K + U,$$

with (from (3)) $\tilde{\beta}_j = \beta_j + \beta_K \gamma_j$. Note that for $j = 0, 1, \ldots, K-1$,

$$\langle Y - \tilde{\beta}_0 - \tilde{\beta}_1 X_1 - \cdots - \tilde{\beta}_{K-1} X_{K-1}, X_j \rangle = \langle \beta_K \tilde{X}_K + U, X_j \rangle = 0.$$

These orthogonality conditions characterize the short linear predictor, and so $\alpha_j = \tilde{\beta}_j$. $\diamond$

The sample counterpart of this result uses the short least-squares fit:

$$\hat{y}_i \,|\, 1, x_1, \ldots, x_{K-1} = a_0 + a_1 x_{i1} + \cdots + a_{K-1} x_{i,K-1}.$$

Claim 4. $a_j = b_j + b_K c_j \qquad (j = 0, 1, \ldots, K-1).$

This least-squares version of the omitted variable bias formula is a computational identity, which can be checked on a data set using a least-squares computer program.

3. MATRIX VERSION OF LINEAR PREDICTOR AND LEAST-SQUARES FIT

Set up the following $(K+1) \times 1$ matrices:

$$X = \begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_K \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}.$$

The linear predictor coefficients β_j are determined by the following orthogonality conditions:

$$\langle Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_K X_K, X_j \rangle = 0 \qquad (j = 0, 1, \ldots, K).$$

So

$$E[(Y - X'\beta)X_j] = E[X_j(Y - X'\beta)] = 0 \qquad (j = 0, 1, \ldots, K).$$

We can write all the orthogonality conditions together as

$$E[X(Y - X'\beta)] = 0.$$

This gives the following system of linear equations:

$$E(XY) - E(XX')\beta = 0,$$

which has the solution

$$\beta = [E(XX')]^{-1} E(XY)$$

(provided that the $(K+1) \times (K+1)$ matrix $E(XX')$ is nonsingular).

For the least-squares fit, set up the $(K+1) \times 1$ matrices

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad x_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad (j = 0, 1, \ldots, K), \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_K \end{pmatrix},$$

and the $n \times (K+1)$ matrix

$$x = \begin{pmatrix} x_0 & x_1 & \dots & x_K \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & \dots & x_{1K} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \dots & x_{nK} \end{pmatrix}.$$

The least-squares coefficients b_j are determined by the following orthogonality conditions:

$$\langle y - b_0 x_0 - b_1 x_1 - \cdots - b_K x_K, x_j \rangle = 0 \qquad (j = 0, 1, \dots, K).$$

So

$$(y - xb)' x_j = x_j'(y - xb) = 0 \qquad (j = 0, 1, \dots, K).$$

We can write all the orthogonality conditions together as

$$\begin{pmatrix} x_0' \\ x_1' \\ \vdots \\ x_K' \end{pmatrix} (y - xb) = x'(y - xb) = 0.$$

This gives the following system of linear equations:

$$x'y - x'xb = 0,$$

which has the solution

$$b = (x'x)^{-1} x'y$$

(provided that the $(K+1) \times (K+1)$ matrix $x'x$ is nonsingular).

LECTURE NOTE 4

PANEL DATA

Consider a population of families. Choose one at random. For each family member t (or for a subset of the family members), there is an outcome variable Y_t and a predictor variable Z_t. There are also variables W and A whose values are the same for all the family members. We have access to data generated by a random sample of size N from this population. The data have realized values of the Y_t, Z_t, and W, but A is not observed. Our objective is to measure a (predictive) effect of Z_t on Y_t, holding constant W and A. We shall develop assumptions that, combined with the family structure of the data, make this feasible.

Consider a population of firms. Choose one at random. For each year from $t = 1, \ldots, T$ there is an output variable Y_t and an input variable Z_t. There is also an input variable A that does not vary over time (but does vary across firms). We have access to data generated by a random sample of size N from this population. The data have realized values of the Y_t and Z_t, but A is not observed. Our objective is to measure a (predictive) effect of Z_t on Y_t holding A constant.

Consider a population of people. Choose a person at random. For each year from $t = 1, \ldots, T$ there is an earnings variable Y_t. There is also a characteristic A of the individual that does not vary over time (but does vary across individuals). We have access to data generated by a random sample of size N from this population. The data have realized values of the Y_t, but A is not observed. Our objective is to measure the (predictive) effect of Y_{t-1} on Y_t holding A constant.

These three examples have much in common, and I shall refer to them as panel data.

The last two examples are a special case called longitudinal data.

1. REGRESSION SYSTEMS

We shall work with the random variables

$$(Y_1, \ldots, Y_T, Z_1, \ldots, Z_T, W, A),$$

which have a joint distribution. For example, we could have a randomly drawn family, from a population of families that have T siblings. The siblings are indexed by $t = 1, \ldots, T$; Y_t is the (adult) earnings of sibling t, Z_t is the education of sibling t, and W is parents' income. The unobserved variable A could be some other measure of family background, such as parents' education.

Consider the regression function for Y_t given $Z_1, \ldots, Z_T, W, A$. We shall assume that Z_s is relevant only for $s = t$:

$$E(Y_t \mid Z_1, \ldots, Z_T, W, A) = g_t(Z_t, W, A),$$

so we have exclusion restrictions on Z_s for $s \neq t$. In addition, we shall impose the following functional form restriction:

$$E(Y_t \mid Z_1, \ldots, Z_T, W, A) = \theta_{0t} + \theta_{1t} Z_t + \theta_{2t} W + \theta_{3t} A,$$

so that the variables enter in a simple linear fashion, with no interactions or higher order polynomial terms. We shall begin with the case in which the coefficients do not depend upon t:

$$E(Y_t \mid Z_1, \ldots, Z_T, W, A) = \theta_0 + \theta_1 Z_t + \theta_2 W + \theta_3 A, \tag{1}$$

and relax that restriction later.

We shall refer to the regression function in equation (1) as a structural regression function. Here "structural" just means that the coefficients in (1) are of direct interest. In particular, θ_1 is the partial (predictive) effect of education on earnings, holding constant

W and A. This regression function does not have a sample counterpart, since A is not observed. So we shall develop linear predictors that do have sample counterparts. Define

$$X' = (\, Z_1 \quad \dots \quad Z_T \quad 1 \quad W \quad X_{T+2} \quad \dots \quad X_K \,),$$

where $X_{T+2}, \dots, X_K$ can include functions (such as polynomials) of the observed variables $Z_1, \dots, Z_T, W$. The linear predictor of Y_t given X is

$$E^*(Y \,|\, X) = \theta_0 + \theta_1 Z_t + \theta_2 W + \theta_3 E^*(A \,|\, X).$$

Our notation for the linear predictor of A given X is

$$E^*(A \,|\, X) = \gamma_1 X_1 + \cdots + \gamma_K X_K.$$

To see how this works, suppose that $T = 2$:

$$E^*(Y_1 \,|\, X) = (\theta_1 + \theta_3\gamma_1) Z_1 + \theta_3\gamma_2 Z_2 + R, \tag{2}$$

$$E^*(Y_2 \,|\, X) = \theta_3\gamma_1 Z_1 + (\theta_1 + \theta_3\gamma_2) Z_2 + R, \tag{3}$$

with

$$R = (\theta_0 + \theta_3\gamma_3) + (\theta_2 + \theta_3\gamma_4) W + \theta_3\gamma_5 X_5 + \cdots + \theta_3\gamma_K X_K.$$

If we look only at the coefficients in the Y_1 predictor, then we cannot identify θ_1. But if we use the Y_1 and Y_2 predictors together, then can obtain θ_1 by subtracting the Z_1 coefficient in predicting Y_2 from the Z_1 coefficient in predicting Y_1. We can also obtain θ_1 by subtracting the Z_2 coefficient in predicting Y_1 from the Z_2 coefficient in predicting Y_2. This is the key to our approach: work with the system of linear predictors for Y_t. (If these were conditional expectation functions, they would be a regression system or multivariate regression.)

The coefficient θ_2 on W in the structural regression function only appears in (2) and (3) in the term $(\theta_2 + \theta_3\gamma_4)$. Since γ_4 does not appear anywhere else, we cannot identify θ_2.

Likewise, we cannot identify θ_0. Also, γ_t only appears multiplied by θ_3. So we can only identify $\theta_3\gamma_1$, $\theta_3\gamma_2$, and $\theta_3\gamma_5, \ldots, \theta_3\gamma_K$.

With general T, we can simplify notation by defining

$$\lambda' \equiv (\, \theta_3\gamma_1 \quad \ldots \quad \theta_3\gamma_T \quad (\theta_0 + \theta_3\gamma_{T+1}) \quad (\theta_2 + \theta_3\gamma_{T+2}) \quad \theta_3\gamma_{T+3} \quad \ldots \quad \theta_3\gamma_K \,).$$

Define

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}, \quad 1_T = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

(1_T is a $T \times 1$ matrix of ones.) Then

$$E^*(Y \,|\, X) = \begin{pmatrix} E^*(Y_1 \,|\, X) \\ \vdots \\ E^*(Y_T \,|\, X) \end{pmatrix} = \Pi X,$$

with

$$\Pi = (\, \theta_1 I_T \quad 0 \,) + 1_T \lambda'. \tag{4}$$

$E^*(Y \,|\, X)$ is our notation for the system of linear predictors (or multivariate linear predictor). The predictor coefficients are arranged in the $T \times K$ matrix Π. (I_T is our notation for the $T \times T$ identity matrix.) Define

$$\alpha' \equiv (\, \theta_1 \quad \lambda_1 \quad \ldots \quad \lambda_T \,).$$

Note that α is unrestricted; there are no restrictions connecting θ_1 and the λ_t. Equation (4) expresses the $T \cdot K$ elements of Π in terms of the $K + 1$ elements of $\alpha' = (\, \theta_1 \quad \lambda' \,)$.

With $T = 3$, we have

$$\Pi = \begin{pmatrix} \theta_1 + \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 & \ldots & \lambda_K \\ \lambda_1 & \theta_1 + \lambda_2 & \lambda_3 & \lambda_4 & \ldots & \lambda_K \\ \lambda_1 & \lambda_2 & \theta_1 + \lambda_3 & \lambda_4 & \ldots & \lambda_K \end{pmatrix}. \tag{5}$$

2. DIFFERENCING TRANSFORMATIONS

We can obtain a simpler system of linear predictors by applying a differencing transformation. The key is a matrix D such that

$$D1_T = 0.$$

Then

$$E^*(DY \mid X) = DE^*(Y \mid X) = D\Pi X = \theta_1 DZ, \tag{6}$$

where

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_T \end{pmatrix}.$$

For example, let D be the $(T-1) \times T$ matrix

$$D = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix}.$$

Then DY and DZ give first differences:

$$DY = \begin{pmatrix} Y_2 - Y_1 \\ \vdots \\ Y_T - Y_{T-1} \end{pmatrix}, \quad DZ = \begin{pmatrix} Z_2 - Z_1 \\ \vdots \\ Z_T - Z_{T-1} \end{pmatrix},$$

and (6) gives

$$E^*(Y_t - Y_{t-1} \mid X) = \theta_1(Z_t - Z_{t-1}) \qquad (t = 2, \dots, T).$$

For a second example, let D be the $T \times T$ matrix

$$D = I_t - \frac{1}{T} 1_T 1_T'.$$

Then DY and DZ give deviations from the means:

$$DY = Y - \bar{Y}1_T = \begin{pmatrix} Y_1 - \bar{Y} \\ \vdots \\ Y_T - \bar{Y} \end{pmatrix}, \quad DZ = Z - \bar{Z}1_T = \begin{pmatrix} Z_1 - \bar{Z} \\ \vdots \\ Z_T - \bar{Z} \end{pmatrix},$$

with $\bar{Y} = \sum_{t=1}^{T} Y_t/T$ and $\bar{Z} = \sum_{t=1}^{T} Z_t/T$. Equation (6) gives

$$E^*(Y_t - \bar{Y} \mid X) = \theta_1(Z_t - \bar{Z}) \qquad (t = 1, \ldots, T).$$

3. IMPOSING RESTRICTIONS

Go back to equation (4). This expresses the $T \cdot K$ elements of Π in terms of the $K+1$ elements of $\alpha' = (\,\theta_1 \quad \lambda'\,)$. So there are restrictions on Π, as we can see in the display in (5). We are going to express the elements of Π as a linear function of α. This leads in the next section to a minimum distance estimator for imposing the restrictions on sample data.

The transpose of Π is the $K \times T$ matrix

$$\Pi' = \theta_1 \begin{pmatrix} I_T \\ 0 \end{pmatrix} + \lambda 1_T' = \theta_1 \,(\, e_1 \quad \ldots \quad e_T \,) + \lambda 1_T',$$

where e_t is a $K \times 1$ matrix of zeros except for a one in row t. Let π be the $K \cdot T \times 1$ matrix formed by stacking the columns of Π':

$$\pi = \text{stack}(\Pi') = \begin{pmatrix} \theta_1 e_1 + \lambda \\ \vdots \\ \theta_1 e_T + \lambda \end{pmatrix}.$$

Now we can express π as a linear function of α:

$$\pi = G \begin{pmatrix} \theta_1 \\ \lambda \end{pmatrix} = G\alpha,$$

where G is the $K \cdot T \times (K+1)$ matrix

$$G = \begin{pmatrix} e_1 & I_K \\ \vdots & \vdots \\ e_T & I_K \end{pmatrix}.$$

The restriction on π is that it is a linear combination of the columns of the known matrix G. So π is restricted to lie in the linear subspace generated by the columns of G; this is a known (given) subspace since G is known (given).

4. MINIMUM DISTANCE ESTIMATION AND GENERALIZED LEAST SQUARES

The data from a sample of $i = 1, \ldots, N$ families can be put in the matrices

$$y_t = \begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix}, \quad z_t = \begin{pmatrix} z_{1t} \\ \vdots \\ z_{Nt} \end{pmatrix}, \quad w = \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} \qquad (t = 1, \ldots, T).$$

We use least squares to form an estimate $\hat{\Pi}$. A least-squares fit of y_t on $z_1, \ldots, z_T, 1, w$ gives the coefficients in row t of Π. (We could also construct data matrices $x_{T+3}, \ldots, x_K$, based on functions of z and w; then these would be included in the least-squares fit.)

Form $\hat{\pi}$ by stacking the columns of $\hat{\Pi}'$:

$$\hat{\pi} = \text{stack}(\hat{\Pi}').$$

Recall that $\alpha' = (\,\theta_1 \quad \lambda'\,)$ and $\pi = G\alpha$. The minimum distance estimate of α is

$$\hat{\alpha} = \arg\min_{\alpha} ||\hat{\pi} - G\alpha||^2.$$

The motivation for this estimator is that the least-squares estimate $\hat{\pi}$ corresponds to the population π but does not impose the restriction that π is a linear combination of the columns of G. So we find the linear combination of the columns of G that gives the best fit to $\hat{\pi}$. The norm in the distance criterion corresponds to the inner product

$$\langle a, b \rangle = a'Cb,$$

where C is a positive definite, symmetric matrix. (A positive definite matrix C is a square matrix, say $J \times J$, such that if a is any nonzero $J \times 1$ matrix, then $a'Ca > 0$; C symmetric means that $C' = C$.) So

$$||\hat{\pi} - G\alpha||^2 = (\hat{\pi} - G\alpha)'C(\hat{\pi} - G\alpha).$$

There is a positive definite, symmetric matrix $C^{1/2}$ that provides a square root of C:

$$C^{1/2}C^{1/2} = C$$

(based on the spectral decomposition of C, from linear algebra). Define

$$\hat{\pi}^* = C^{1/2}\hat{\pi}, \quad G^* = C^{1/2}G.$$

Then $\hat{\alpha}$ can be obtained from a least-squares fit of $\hat{\pi}^*$ on G^* (and we use the matrix version from Section 3 of Note 3):

$$\begin{aligned}\hat{\alpha} &= \arg\min_{\alpha}(\hat{\pi}^* - G^*\alpha)'(\hat{\pi}^* - G^*\alpha) \\ &= (G^{*\prime}G^*)^{-1}G^{*\prime}\hat{\pi}^* \\ &= (G'CG)^{-1}G'C\hat{\pi}. \end{aligned} \tag{7}$$

The expression for $\hat{\alpha}$ in (7) is known as *generalized least squares*.

We are free to choose the weight matrix C, and this makes the minimum distance estimator more flexible. If some components of $\hat{\pi}$ are estimated more precisely than others, then we may want to give more weight to those components. We will discuss an optimal choice for C in the inference part of the course, where we work out the distribution of $\hat{\pi}$ in repeated samples. For now, we can just set C equal to an identity matrix.

5. STACKING

There is another way to impose restrictions, in which the data matrices are stacked and then used in a least-squares fit. I'll use the difference transformations from Section 2 to illustrate. Define

$$\tilde{Y}_t = Y_t - Y_{t-1}, \quad \tilde{Z}_t = Z_t - Z_{t-1} \qquad (t = 2, \ldots, T).$$

Set up the corresponding data matrices

$$\tilde{y}_t = \begin{pmatrix} y_{1t} - y_{1,t-1} \\ \vdots \\ y_{Nt} - y_{N,t-1} \end{pmatrix}, \quad \tilde{z}_t = \begin{pmatrix} z_{1t} - z_{1,t-1} \\ \vdots \\ z_{Nt} - z_{N,t-1} \end{pmatrix} \qquad (t = 2, \ldots, T).$$

We have

$$E^*(\tilde{Y}_t \,|\, \tilde{Z}_t) = \pi_t \tilde{Z}_t,$$

with $\pi_t = \theta_1$ $(t = 2, \ldots, T)$. The sample counterpart is to obtain an estimate $\hat{\pi}_t$ from the least-squares fit of $\tilde{y}_t$ on $\tilde{z}_t$. Define

$$\pi = \begin{pmatrix} \pi_2 \\ \vdots \\ \pi_T \end{pmatrix}, \quad \hat{\pi} = \begin{pmatrix} \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_T \end{pmatrix}.$$

Then

$$\pi = G\theta_1 \quad \text{with} \quad G = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1_{T-1},$$

and we can obtain a minimum distance estimate of θ_1 from

$$\hat{\theta}_1 = \arg\min_{\theta_1} ||\hat{\pi} - G\theta_1||^2 = (G'CG)^{-1}G'C\hat{\pi}. \tag{8}$$

Now stack up the data matrices as follows:

$$\tilde{y} = \begin{pmatrix} \tilde{y}_2 \\ \vdots \\ \tilde{y}_T \end{pmatrix}, \quad \tilde{z} = \begin{pmatrix} \tilde{z}_2 \\ \vdots \\ \tilde{z}_T \end{pmatrix}.$$

We can obtain an estimate of θ_1 from a least-squares fit of $\tilde{y}$ on $\tilde{z}$:

$$\hat{\theta}_1 = (\tilde{z}'\tilde{z})^{-1}\tilde{z}'\tilde{y} = \left(\sum_{t=2}^{T} \tilde{z}_t'\tilde{z}_t\right)^{-1} \sum_{t=2}^{T} \tilde{z}_t'\tilde{y}_t. \tag{9}$$

This stacked estimate equals the minimum distance estimate for a particular choice of the weight matrix C. Let

$$C = \text{diag}(\tilde{z}_2'\tilde{z}_2, \ldots, \tilde{z}_T'\tilde{z}_T) = \begin{pmatrix} \tilde{z}_2'\tilde{z}_2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \tilde{z}_T'\tilde{z}_T \end{pmatrix}.$$

Because $\hat{\pi}_t = (\tilde{z}_t'\tilde{z}_t)^{-1}\tilde{z}_t'\tilde{y}_t$, we have

$$(G'CG)^{-1}G'C\hat{\pi} = \left(\sum_{t=2}^{T} \tilde{z}_t'\tilde{z}_t\right)^{-1} \sum_{t=2}^{T} \tilde{z}_t'\tilde{y}_t.$$

So for this choice of the weight matrix C, the minimum distance estimate in (8) equals the stacked estimate in (9).

Similar points apply with the deviations from means transformation. Now let

$$\tilde{Y}_t = Y_t - \bar{Y}, \quad \tilde{Z}_t = Z_t - \bar{Z} \qquad (t = 1, \ldots, T),$$

and set up the corresponding data matrices

$$\tilde{y}_t = \begin{pmatrix} y_{1t} - \bar{y}_1 \\ \vdots \\ y_{Nt} - \bar{y}_N \end{pmatrix}, \quad \tilde{z}_t = \begin{pmatrix} z_{1t} - \bar{z}_1 \\ \vdots \\ z_{Nt} - \bar{z}_N \end{pmatrix},$$

with

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it} \qquad (i = 1, \ldots, N).$$

We have

$$E^*(\tilde{Y}_t \,|\, \tilde{Z}_t) = \pi_t \tilde{Z}_t,$$

with $\pi_t = \theta_1$ $(t = 1, \ldots, T)$. The sample counterpart is to obtain an estimate $\hat{\pi}_t$ from the least-squares fit of $\tilde{y}_t$ on $\tilde{z}_t$. Define

$$\hat{\pi} = \begin{pmatrix} \hat{\pi}_1 \\ \vdots \\ \hat{\pi}_T \end{pmatrix}.$$

Stack up the data matrices:

$$\tilde{y} = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_T \end{pmatrix}, \quad \tilde{z} = \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_T \end{pmatrix}.$$

We can obtain an estimate of θ_1 from a least-squares fit of $\tilde{y}$ on $\tilde{z}$:

$$\hat{\theta}_1 = (\tilde{z}'\tilde{z})^{-1}\tilde{z}'\tilde{y} = \left(\sum_{t=1}^{T} \tilde{z}_t'\tilde{z}_t\right)^{-1} \sum_{t=1}^{T} \tilde{z}_t'\tilde{y}_t. \tag{10}$$

This stacked estimate equals the minimum distance estimate for a particular choice of the weight matrix C. Let

$$C = \text{diag}(\tilde{z}_1'\tilde{z}_1, \ldots, \tilde{z}_T'\tilde{z}_T) = \begin{pmatrix} \tilde{z}_1'\tilde{z}_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \tilde{z}_T'\tilde{z}_T \end{pmatrix}.$$

Because $\hat{\pi}_t = (\tilde{z}_t'\tilde{z}_t)^{-1}\tilde{z}_t'\tilde{y}_t$, we have

$$(G'CG)^{-1}G'C\hat{\pi} = \left(\sum_{t=1}^{T} \tilde{z}_t'\tilde{z}_t\right)^{-1} \sum_{t=1}^{T} \tilde{z}_t'\tilde{y}_t.$$

So for this choice of the weight matrix C, the minimum distance estimate equals the stacked estimate in (10).

In general, the stacked estimate of θ_1 based on the first difference transformation is not equal to the stacked estimate based on the deviations from means transformation. (They are equal if $T = 2$.) The choice between these estimates depends on their sampling distributions and will be discussed in the inference part of the course.

6. PRODUCTION FUNCTION

For a randomly chosen farm, Q_t is output in year t, L_t is labor input in year t, F is a measure of soil quality and other location aspects that are not changing over time, and V_t is a measure of rainfall and other weather conditions in year t. The production function is

$$Q_t = L_t^{\theta} F V_t \qquad (t = 1, \ldots, T)$$

with $0 < \theta < 1$. Data are available on output and labor input for N of these farms over T years; data on soil quality and weather conditions are not available. The farmer's objective is to choose the labor input to maximize the conditional expectation of profit, conditional on the information $\mathcal{J}_t$ available to him when the labor choice is made:

$$\max_{L} E[P_t Q_t - W_t L \,|\, \mathcal{J}_t].$$

The price of output (P_t) and the price of labor (W_t) are not affected by the farmer's choice. The first-order condition for

$$\max_L P_t[L^\theta F E(V_t \mid \mathcal{J}_t)] - W_t L$$

gives

$$\theta P_t L^{\theta-1} F E(V_t \mid J_t) = W_t.$$

The derived demand for labor is

$$\log L_t = \frac{1}{1-\theta}[\log\theta - \log\frac{W_t}{P_t} + \log F + \log E(V_t \mid \mathcal{J}_t)].$$

We can write the production function as

$$\log Q_t = \theta \log L_t + \log F + \log V_t,$$

or

$$Y_t = \theta Z_t + A + \log V_t,$$

with $Y_t = \log Q_t$, $Z_t = \log L_t$, and $A = \log F$. Note that A is correlated with Z_t through the derived demand for labor, and so a regression function for Y_t that does not include A will not have the production function elasticity θ as the coefficient on Z_t. This is the omitted variable bias motivation for the use of panel data.

Note that

$$E(Y_t \mid Z_1, \ldots, Z_T, A) = \theta Z_t + A + E(\log V_t \mid Z_1, \ldots, Z_T, A).$$

Our structural regression function in equation (1) in Section 1 has the form

$$E(Y_t \mid Z_1, \ldots, Z_T, A) = \theta Z_t + A + \text{constant}.$$

So to apply the results from Section 1, with θ interpreted as the production function elasticity, we need

$$E(\log V_t \mid Z_1, \ldots, Z_T, A) = \text{constant}.$$

This could fail to hold if there is correlation over time in the weather conditions V_t. For then lagged values of $\log V_t$ will enter the demand for labor, and so Z_{t+1} will be correlated with $\log V_t$.

7. TIME-VARYING COEFFICIENTS

Return to the structural regression function in equation (1) of Section 1, and allow for time-varying coefficients:

$$E(Y_t \mid Z_1, \ldots, Z_T, W, A) = \theta_{0t} + \theta_{1t} Z_t + \theta_{2t} W + \theta_{3t} A.$$

As before, let

$$X' = (\, Z_1 \quad \ldots \quad Z_T \quad 1 \quad W \quad X_{T+3} \quad \ldots \quad X_K \,),$$

$$E^*(A \mid X) = \gamma_1 X_1 + \cdots + \gamma_K X_K = \gamma' X.$$

Then the linear predictor of Y_t given X is

$$E^*(Y_t \mid X) = \theta_{0t} + \theta_{1t} Z_t + \theta_{2t} W + \theta_{3t} \gamma' X.$$

The multivariate linear predictor is

$$E^*(Y \mid X) = \Pi X \quad \text{with} \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}.$$

The matrix Π of linear predictor coefficients is $T \times K$. With $T = 3$,

$$\Pi = \begin{pmatrix} \theta_{11} + \theta_{31}\gamma_1 & \theta_{31}\gamma_2 & \theta_{31}\gamma_3 & \ldots \\ \theta_{32}\gamma_1 & \theta_{12} + \theta_{32}\gamma_2 & \theta_{32}\gamma_3 & \ldots \\ \theta_{33}\gamma_1 & \theta_{33}\gamma_2 & \theta_{13} + \theta_{33}\gamma_3 & \ldots \end{pmatrix}$$

(where we are displaying only the first three columns of Π). Note that

$$\frac{\pi_{12}}{\pi_{32}} = \frac{\theta_{31}\gamma_2}{\theta_{33}\gamma_2} = \frac{\theta_{31}}{\theta_{33}},$$

$$\pi_{31} = \theta_{33}\gamma_1,$$

and so

$$\pi_{11} - \frac{\pi_{12}}{\pi_{32}}\pi_{31} = \theta_{11}.$$

So θ_{11} is identified (provided that $\pi_{32} \neq 0$), and a similar argument shows that θ_{1t} is identified for $t = 1, 2, 3$.

LECTURE NOTE 5

AUTOREGRESSION IN PANEL DATA

1. STRUCTURAL REGRESSION MODEL

Consider a population of people. Choose a person at random. For each year from $t = 1, \ldots, T$ there is an earnings variable Y_t. There is also a characteristic A of the individual that does not vary over time (but does vary across individuals). We have access to data generated by a random sample of size N from this population. The data have realized values of the Y_t, but A is not observed. Our objective is to measure the (predictive) effect of Y_{t-1} on Y_t holding A constant.

The structural regression model is

$$E(Y_t \,|\, Y_1, \ldots, Y_{t-1}, A) = \lambda_t + \theta Y_{t-1} + A \qquad (t = 2, \ldots, T). \tag{1}$$

This is called autoregression because we are predicting Y_t using the past values of the same variable. Two sorts of restrictions are being imposed. We condition on all the past values from Y_1 to Y_{t-1}, but the assumption in (1) is that only Y_{t-1} matters (once we control for A). So there are exclusion restrictions. The other sort of restriction in (1) is that the functional form is very simple, with no interaction terms involving Y_{t-1} and A. The intercept λ_t allows for an unrestricted additive period effect. The partial effects of Y_{t-1} and A are assumed to be constant over time (but we could extend the analysis and allow for $\theta_{1t} Y_{t-1} + \theta_{2t} A$). Given the assumption that the partial effect of A is constant over time, it is not restrictive to set the coefficient on A equal to one. Because A is not observed, we can scale it so that the coefficient is one.

The sample data are y_{it} for $i = 1, \ldots, N$ individuals and $t = 1, \ldots, T$ periods. For

example, y_{it} could be the earnings of person i in year t. Data on the variable A are not available.

2. SOLVING THE MODEL

We are going to solve out the lagged dependent variables and express Y_t as a function just of A and prediction errors. This will lead to a formula for the covariances between Y_t and Y_s $(s, t = 1, \ldots, T)$. We will see that θ can be expressed as a function of these covariances. Then we can use the sample covariances to obtain an estimate of θ.

To solve the model, we must do something about Y_1, because we no not observe Y_0. We handle this initial conditions problem by introducing a linear predictor for Y_1:

$$E^*(Y_1 \,|\, 1, A) = \delta_0 + \delta_1 A.$$

Note that this does not introduce additional restrictions. All the restrictions are expressed in the structural regression model in (1). We shall impose the normalization that

$$E(A) = 0.$$

This is not restrictive, because we could redefine A as $A - E(A)$, and redefine the period effect λ_t as $\lambda_t + E(A)$. The period effects would still be unrestricted.

Define the prediction errors

$$U_t = Y_t - E^*(Y_t \,|\, 1, Y_1, \ldots, Y_{t-1}, A) \qquad (t = 2, \ldots, T),$$
$$V = Y_1 - E^*(Y_1 \,|\, 1, A).$$

Then we have

$$Y_1 = \delta_0 + \delta_1 A + V, \tag{2}$$

$$Y_t = \lambda_t + \theta Y_{t-1} + A + U_t \qquad (t = 2, \ldots, T). \tag{3}$$

By construction, the prediction error V is orthogonal to 1 and A. So $E(V) = 0$ and V is uncorrelated with A. Likewise, by construction, the prediction error U_t is orthogonal

to $1, Y_1, \ldots, Y_{t-1}, A$. So $E(U_t) = 0$ and U_t is uncorrelated with $Y_1, \ldots, Y_{t-1}$, and with A. Because U_{t-1} is a linear function of Y_{t-1}, Y_{t-2}, and A, the covariance between U_t and U_{t-1} is zero. Likewise, U_t is uncorrelated with $U_{t-2}, \ldots, U_2$. In addition, since V is a linear function of Y_1 and A, the covariance between U_t and V is zero. So after we have solved out the lagged Y's, and have expressed the Y's as linear functions of $A, U_2, \ldots, U_T, V$, it will be straightforward to obtain a formula for the covariances between the Y's.

Substitute for Y_1 in the equation for Y_2 in (3):

$$\begin{aligned} Y_2 &= \lambda_2 + \theta[\delta_0 + \delta_1 A + V] + A + U_2 \\ &= (\lambda_2 + \theta\delta_0) + (1 + \theta\delta_1)A + \theta V + U_2. \end{aligned} \tag{4}$$

Recursively substitute for Y_2 from (4) into the equation for Y_3 in (3):

$$\begin{aligned} Y_3 &= \lambda_3 + \theta[(\lambda_2 + \theta\delta_0) + (1 + \theta\delta_1)A + \theta V + U_2] + A + U_3 \\ &= (\lambda_3 + \theta\lambda_2 + \theta^2\delta_0) + (1 + \theta + \theta^2\delta_1)A + \theta^2 V + U_3 + \theta U_2. \end{aligned}$$

We could continue with this recursive substitution, but it is more convenient to use matrix notation. With $T = 3$:

$$\begin{pmatrix} 1 & 0 & 0 \\ -\theta & 1 & 0 \\ 0 & -\theta & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} \delta_0 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ 1 \\ 1 \end{pmatrix} A + \begin{pmatrix} V \\ U_2 \\ U_3 \end{pmatrix}.$$

We can write this as

$$B(\theta)Y = \begin{pmatrix} \delta_0 \\ \lambda \end{pmatrix} + \begin{pmatrix} \delta_1 \\ 1_2 \end{pmatrix} A + \begin{pmatrix} V \\ U \end{pmatrix}.$$

It is straightforward to check that

$$B(\theta)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \theta & 1 & 0 \\ \theta^2 & \theta & 1 \end{pmatrix}.$$

The recursive substitution to solve out the lagged Y's corresponds to multiplying by $B(\theta)^{-1}$:

$$Y = B(\theta)^{-1} \begin{pmatrix} \delta_0 \\ \lambda \end{pmatrix} + B(\theta)^{-1} \left[\begin{pmatrix} \delta_1 \\ 1_2 \end{pmatrix} A + \begin{pmatrix} V \\ U \end{pmatrix} \right]. \tag{5}$$

3. COVARIANCE MATRIX

With

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}$$

a $T \times 1$ matrix of random variables, define $E(Y)$ to be the $T \times 1$ matrix with t^{th} element equal to $E(Y_t)$. Define $\text{Cov}(Y)$ to be the $T \times T$ matrix with (s,t) element equal to $\text{Cov}(Y_s, Y_t)$. For notation, let $\mu = E(Y)$ and $\Sigma = \text{Cov}(Y)$. With $T = 3$:

$$\mu = E(Y) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ E(Y_3) \end{pmatrix},$$

$$\Sigma = \text{Cov}(Y) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \quad \text{with} \quad \sigma_{st} = \text{Cov}(Y_s, Y_t) = \sigma_{ts}.$$

We shall focus on the population $E(Y)$ and covariance matrix $\text{Cov}(Y)$ because they have direct sample counterparts. Given the data y_{it} $(i = 1, \ldots, N;\ t = 1, \ldots, T)$, we can estimate the population means using sample means:

$$\hat{\mu}_t = \bar{y}_t = \frac{1}{N} \sum_{t=1}^{T} y_{it}.$$

Note that

$$\sigma_{st} = \text{Cov}(Y_s, Y_t) = E[(Y_s - \mu_s)(Y_t - \mu_t)] = E(Y_s Y_t) - \mu_s \mu_t.$$

This suggests the estimate

$$\hat{\sigma}_{st} = \frac{1}{N} \sum_{i=1}^{N} (y_{is} - \bar{y}_s)(y_{it} - \bar{y}_t) = \frac{1}{N} \sum_{i=1}^{N} y_{is} y_{it} - \bar{y}_s \bar{y}_t. \tag{6}$$

The following result makes it easy to use equation (5) express the covariance matrix of Y as an explicit function of θ and some other parameters.

Claim 1. Suppose that the random matrix Y is $T \times 1$; and the nonrandom matrices d_1 and d_2 are $M \times T$ and $M \times 1$. Then

$$\text{Cov}(d_1 Y + d_2) = d_1 \text{Cov}(Y) d_1'.$$

Proof. First note that with $\tilde{Y} = Y - E(Y)$,

$$\text{Cov}(Y) = E(\tilde{Y}\tilde{Y}').$$

With $W = d_1 Y + d_2$, we have

$$\begin{aligned} E(W) &= d_1 E(Y) + d_2 \\ \tilde{W} &= W - E(W) = d_1[Y - E(Y)] = d_1 \tilde{Y}, \end{aligned}$$

and so

$$\text{Cov}(W) = E(\tilde{W}\tilde{W}') = d_1 E(\tilde{Y}\tilde{Y}') d_1' = d_1 \text{Cov}(Y) d_1'. \quad \diamond$$

Applying this result to equation (5) (with $T = 3$) gives

$$\Sigma = \text{Cov}(Y) = B(\theta)^{-1} \left[\begin{pmatrix} \delta_1 \\ 1_2 \end{pmatrix} \sigma_A^2 \begin{pmatrix} \delta_1 & 1_2' \end{pmatrix} + \begin{pmatrix} \sigma_v^2 & 0 & 0 \\ 0 & \sigma_{u_2}^2 & 0 \\ 0 & 0 & \sigma_{u_3}^2 \end{pmatrix} \right] B(\theta)^{-1\prime}, \qquad (7)$$

with $\sigma_A^2 = \text{Var}(A)$, $\sigma_v^2 = \text{Var}(V)$, and $\sigma_{u_t}^2 = \text{Var}(U_t)$. With general T, we have

$$\Sigma = \text{Cov}(Y) = B(\theta)^{-1} \left[\begin{pmatrix} \delta_1 \\ 1_{T-1} \end{pmatrix} \begin{pmatrix} \delta_1 & 1_{T-1}' \end{pmatrix} \sigma_A^2 + \text{diag}(\sigma_v^2, \sigma_{u_2}^2, \ldots, \sigma_{u_T}^2) \right] B(\theta)^{-1\prime}, \quad (8)$$

with

$$B(\theta)^{-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \theta & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta^{T-1} & \theta^{T-2} & \ldots & 1 \end{pmatrix}.$$

With $T = 3$, let σ be a 6×1 matrix formed from the distinct elements of the symmetric 3×3 matrix Σ:

$$\sigma' = (\sigma_{11} \quad \sigma_{21} \quad \sigma_{31} \quad \sigma_{22} \quad \sigma_{32} \quad \sigma_{33}) .$$

Define the 6×1 matrix of parameters α:

$$\alpha' = (\theta \quad \delta_1 \quad \sigma_A^2 \quad \sigma_v^2 \quad \sigma_{u_2}^2 \quad \sigma_{u_3}^2) .$$

Equation (7) allows us to express σ as a known function $g(\cdot)$ of α:

$$\sigma = g(\alpha).$$

Use the sample covariances $\hat{\sigma}_{st}$ in (6) to obtain the estimate $\hat{\sigma}$. Then obtain an estimate $\hat{\alpha}$ by solving

$$\hat{\alpha} = \arg\min_{\alpha} ||\hat{\sigma} - g(\alpha)||^2.$$

Since $\hat{\sigma}$ are α are both 6×1, we expect to be able to find an $\hat{\alpha}$ that gives a perfect fit: $\hat{\sigma} = g(\hat{\alpha})$.

With general T, let σ be a $T(T+1)/2 \times 1$ matrix formed from the distinct elements of the symmetric $T \times T$ matrix Σ:

$$\sigma' = (\sigma_{11} \quad \ldots \quad \sigma_{T1} \quad \sigma_{22} \quad \ldots \quad \sigma_{T2} \quad \ldots \quad \sigma_{TT}) .$$

Define the $(T+3) \times 1$ matrix of parameters α:

$$\alpha' = (\theta \quad \delta_1 \quad \sigma_A^2 \quad \sigma_v^2 \quad \sigma_{u_2}^2 \quad \ldots \quad \sigma_{u_T}^2) .$$

Equation (8) allows us to express σ as a known function $g(\cdot)$ of α:

$$\sigma = g(\alpha). \tag{9}$$

Note that the parameters in α are unrestricted. When $T = 4$, there are 10 parameters in σ and only 7 parameters in α. So the structural regression model in equation (1) imposes

restrictions on σ, and the minimum distance estimator provides a method for imposing those restrictions. As in Section 4 of Note 4, we are free to choose a positive definite, symmetric weight matrix C:

$$\hat{\alpha} = \arg\min_{\alpha} ||\hat{\sigma} - g(\alpha)||^2 = \arg\min_{\alpha} (\hat{\sigma} - g(\alpha))' C (\hat{\sigma} - g(\alpha)). \tag{10}$$

Later in the course we shall discuss an optimal choice for C; for now, we can use an identity matrix.

Once we have obtained an estimate of θ from $\hat{\alpha}$, we can use

$$E(Y) = B(\theta)^{-1} \begin{pmatrix} \delta_0 \\ \lambda_2 \\ \vdots \\ \lambda_T \end{pmatrix}$$

to obtain estimates of δ_0 and the λ_t:

$$\begin{pmatrix} \hat{\delta}_0 \\ \hat{\lambda}_2 \\ \vdots \\ \hat{\lambda}_T \end{pmatrix} = B(\hat{\theta}) \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_T \end{pmatrix}.$$

5. IDENTIFICATION AND EXPLICIT SOLUTIONS

We shall say that θ is identified if we can solve for θ given the population distribution of observed variables. The population distribution of Y determines $\Sigma = \text{Cov}(Y)$, and so θ is identified if we can solve for θ from Σ.

Consider the differencing transformation:

$$Y_3 - Y_2 = \lambda_3 - \lambda_2 + \theta(Y_2 - Y_1) + U_3 - U_2.$$

This implies that

$$\text{Cov}(Y_3 - Y_2 - \theta(Y_2 - Y_1), Y_1) = \text{Cov}(\lambda_3 - \lambda_2 + U_3 - U_2, Y_1) = 0. \tag{11}$$

We are using

$$\mathrm{Cov}(U_3, Y_1) = \mathrm{Cov}(U_2, Y_1) = 0,$$

which follows from the definition of U_t as a prediction error. We can obtain from (11) an equation involving Σ and θ:

$$\sigma_{31} - \sigma_{21} - \theta(\sigma_{21} - \sigma_{11}) = 0.$$

So

$$\theta = \frac{\sigma_{31} - \sigma_{21}}{\sigma_{21} - \sigma_{11}}, \tag{12}$$

provided that $\sigma_{21} \neq \sigma_{11}$. Equation (12) suggests the following explicit estimator for θ:

$$\hat{\theta} = \frac{\hat{\sigma}_{31} - \hat{\sigma}_{21}}{\hat{\sigma}_{21} - \hat{\sigma}_{11}}.$$

For general T, the differencing transformation gives

$$Y_t - Y_{t-1} = \lambda_t - \lambda_{t-1} + \theta(Y_{t-1} - Y_{t-2}) + U_t - U_{t-1},$$

which implies that

$$\mathrm{Cov}(Y_t - Y_{t-1} - \theta(Y_{t-1} - Y_{t-2}), Y_{t-j}) = \mathrm{Cov}(\lambda_t - \lambda_{t-1} + U_t - U_{t-1}, Y_{t-j}) = 0$$

if $j \geq 2$. So we have

$$\sigma_{t,t-j} - \sigma_{t-1,t-j} - \theta(\sigma_{t-1,t-j} - \sigma_{t-2,t-j}) = 0,$$

and

$$\theta = \frac{\sigma_{t,t-j} - \sigma_{t-1,t-j}}{\sigma_{t-1,t-j} - \sigma_{t-2,t-j}} \qquad (t = 3, \ldots, T;\ j = 2, \ldots, t-1). \tag{13}$$

With $T = 4$, equation (13) gives three solutions for θ, from $j = 2$ with $t = 3$ and $j = 2, 3$ with $t = 4$.

There are additional solutions for θ that involve quadratic equations. Because

$$Y_t - \theta Y_{t-1} = \lambda_t + A + U_t,$$

we have

$$\mathrm{Cov}(Y_t - \theta Y_{t-1}, Y_s - \theta Y_{s-1}) = \sigma_A^2$$

if $t \neq s$. So

$$\mathrm{Cov}(Y_3 - \theta Y_2, Y_2 - \theta Y_1) = \sigma_A^2 = \mathrm{Cov}(Y_4 - \theta Y_3, Y_3 - \theta Y_2).$$

This implies that

$$\sigma_{32} - \theta\sigma_{22} - \theta\sigma_{31} + \theta^2\sigma_{21} = \sigma_{43} - \theta\sigma_{33} - \theta\sigma_{42} + \theta^2\sigma_{32},$$

which gives a quadratic equation to solve for θ.

An advantage of the minimum distance estimator in (10) is that it imposes all of the restrictions on Σ and so makes use of all the ways in which we can solve for θ from Σ.

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

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

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 $x = g(z)$:

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

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

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

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

$$r(w) = E(Y_i \mid Z_i = w).$$

(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

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

and

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

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 $k = 1, \ldots, K$). 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

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

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

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

The off-diagonal terms are

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

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* $s(\cdot)$:

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

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

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

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

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

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

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

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

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

$\operatorname{Cov}(b \mid Z = z)$ is $K \times K$:

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

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

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

We can summarize our results as follows.

Claim 1. Under random sampling, if

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

then

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

LECTURE NOTE 7

NORMAL LINEAR MODEL

1. NORMAL DISTRIBUTION

We shall use $\mathcal{N}(0,1)$ to denote the standard normal distribution. It is a continuous distribution with a density function

$$f(w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}w^2).$$

Probabilities are obtained by integrating the density: if the random variable W has a $\mathcal{N}(0,1)$ distribution, then

$$\text{Prob}(c \leq W \leq d) = \int_c^d f(w)\, dw.$$

The expectation of W is zero and the variance is one: $E(W) = 0$, $\text{Var}(W) = 1$.

A general normal distribution, with mean μ and variance σ^2, is obtained from a linear function of a standard normal: if $W \sim \mathcal{N}(0,1)$, then

$$\mu + \sigma W \ \sim \ \mathcal{N}(\mu, \sigma^2).$$

2. NORMAL LINEAR MODEL

We continue to assume random sampling:

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

In this note, Y_i is 1×1 ($M = 1$). We make the two assumptions in Claim 1 of Note 6:

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

In addition, we assume that the conditional distribution of Y_i given $Z_i = z_i$ is normal:

$$Y_i \,|\, Z_i = z_i \;\sim\; \mathcal{N}(x_i'\beta, \sigma^2) \qquad (i = 1, \ldots, n).$$

Define the prediction error

$$U_i \equiv Y_i - x_i'\beta,$$

and let

$$V_i = U_i/\sigma,$$

so that

$$V_i \,|\, Z_i = z_i \;\sim\; \mathcal{N}(0, 1).$$

Because of the random sampling, this implies that

$$V_i \,|\, Z = z \;\overset{\text{i.i.d.}}{\sim}\; \mathcal{N}(0, 1) \qquad (i = 1, \ldots, n).$$

If W is $n \times 1$ with a joint normal distribution, with mean μ and covariance matrix Σ, we shall use the notation $W \sim \mathcal{N}(\mu, \Sigma)$. So we can write the normal linear model as

$$Y = x\beta + \sigma V, \quad V \,|\, Z = z \;\sim\; \mathcal{N}(0, I_n), \tag{1}$$

with

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad x = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}, \quad V = \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix}$$

(and $x = g(z)$.) We are going to transform this model into the following canonical form:

$$Y^* = \begin{pmatrix} s \\ 0 \end{pmatrix} \beta + \sigma V^*, \qquad V^* \,|\, Z = z \;\sim\; \mathcal{N}(0, I_n). \tag{2}$$

This will make it simpler to obtain the distribution of the sum of squared residuals (from a least-squares fit), and to show that the least-squares estimator is independent of the sum of squared residuals.

3. CANONICAL FORM

A $n \times n$ matrix q is *orthogonal* if $q^{-1} = q'$. Note that if q is orthogonal then so is q', and $qq' = q'q = I_n$. Orthogonal matrices preserve the least-squares inner product:

Claim 1. If c and d are $n \times 1$ and q is an orthogonal $n \times n$ matrix, then

$$\langle qc, qd \rangle = \langle c, d \rangle.$$

Proof.

$$\langle qc, qd \rangle = (qc)'(qd) = c'q'qd = c'd = \langle c, d \rangle. \quad \diamond$$

So an orthogonal matrix preserves the least-squares norm:

$$||qc||^2 = \langle qc, qc \rangle = \langle c, c \rangle = ||c||^2.$$

Claim 2 (QR decomposition). If x is $n \times K$, then there is an orthogonal $n \times n$ matrix q and an upper triangular $n \times K$ matrix r such that

$$x = qr.$$

If $n > K$, then

$$r = \begin{pmatrix} s \\ 0 \end{pmatrix}, \quad s = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1K} \\ 0 & s_{22} & \dots & s_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_{KK} \end{pmatrix}.$$

See G. Golub and C. Van Loan, *Matrix Computations*, Third Edition, The Johns Hopkins University Press, 1996.

We shall use the following property of joint normality:

Claim 3. If d_1 is $m \times n$, d_2 is $m \times 1$, with d_1 and d_2 not random, and if W is $n \times 1$ with $W \sim \mathcal{N}(\mu, \Sigma)$, then

$$d_1 W + d_2 \sim \mathcal{N}(d_1\mu + d_2, d_1 \Sigma d_1').$$

Claim 4. If $V \sim \mathcal{N}(0, I_n)$ and q is an $n \times n$ orthogonal (nonrandom) matrix, then

$$qV \sim \mathcal{N}(0, I_n).$$

Proof.

$$qV \sim \mathcal{N}(0, qI_n q') = \mathcal{N}(0, I_n).$$

Now we can transform the normal linear model in (1) into the canonical form in (2). We shall assume that $n > K$ and that there is no exact linear dependence connecting the columns of x: if c is $K \times 1$ and $xc = 0$, then $c = 0$. This implies that $x'x$ is nonsingular and so s in the QR decomposition of x is nonsingular. From (1),

$$Y = q \begin{pmatrix} s \\ 0 \end{pmatrix} \beta + \sigma V.$$

Multiply both sides of this equation by q':

$$q'Y = \begin{pmatrix} s \\ 0 \end{pmatrix} \beta + \sigma q'V.$$

Define

$$Y^* = q'Y, \quad V^* = q'V,$$

and use Claim 4 to obtain our canonical form:

$$Y^* = \begin{pmatrix} s \\ 0 \end{pmatrix} \beta + \sigma V^*, \qquad V^* \mid Z = z \sim \mathcal{N}(0, I_n).$$

Define

$$Y^*_{(1)} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_K \end{pmatrix}, \quad V^*_{(1)} = \begin{pmatrix} V_1 \\ \vdots \\ V_K \end{pmatrix},$$

and

$$Y^*_{(2)} = \begin{pmatrix} Y_{K+1} \\ \vdots \\ Y_n \end{pmatrix}, \quad V^*_{(2)} = \begin{pmatrix} V_{K+1} \\ \vdots \\ V_n \end{pmatrix}.$$

Then we can write the canonical form as

$$Y^*_{(1)} = s\beta + \sigma V^*_{(1)}, \tag{3}$$

$$Y^*_{(2)} = \sigma V^*_{(2)}, \tag{4}$$

where s is a $K \times K$ nonsingular matrix and, conditional on $Z = z$, the components of $V^*_{(1)}$ and $V^*_{(2)}$ are all i.i.d. $\mathcal{N}(0,1)$.

The least squares estimator solves

$$b = \arg\min_d ||Y - xd||^2.$$

From Claim 1, and using the QR decomposition of x,

$$\begin{aligned} ||Y - xd||^2 = ||q'(Y - xd)||^2 &= ||Y^* - \begin{pmatrix} s \\ 0 \end{pmatrix} d||^2 \\ &= ||Y^*_{(1)} - sd||^2 + ||Y^*_{(2)}||^2. \end{aligned}$$

So we have

$$b = s^{-1}Y^*_{(1)} = \beta + \sigma s^{-1}V^*_{(1)}, \tag{5}$$

and the sum of squared residuals is

$$\begin{aligned} \text{SSR} \equiv \min_d ||Y - xd||^2 &= ||Y - xb||^2 \\ &= ||Y_{(2)}||^2 = \sigma^2 ||V_{(2)}||^2 \\ &= \sigma^2 \sum_{i=K+1}^{n} (V^*_i)^2. \end{aligned} \tag{6}$$

Since the least-squares estimator depends on V^*_i only for $i = 1, \ldots, K$, and since SSR depends on V^*_i only for $i = K+1, \ldots, n$, we have b independent of SSR conditional on $Z = z$.

From (5), the distribution of the least-squares estimator is joint normal:

$$b \,|\, Z = z \sim \mathcal{N}(\beta, \sigma^2 s^{-1} s^{-1\prime}) = \mathcal{N}(\beta, \sigma^2 (x'x)^{-1}).$$

Definition 1. If V_i i.i.d $\mathcal{N}(0,1)$, then

$$\sum_{i=1}^{m} V_i^2 \sim \text{Chi}^2(m).$$

We shall also use the notation $\chi^2(m)$. The parameter m is called the degrees of freedom. Note that the mean of a random variable with a chi-square distribution is

$$E[\text{Chi}^2(m)] = E(\sum_{i=1}^{m} V_i^2) = m.$$

From (6), the sum of squared residuals is distributed as σ^2 times a random variable with a chi-square distribution, with $n-K$ degrees of freedom:

$$\text{SSR} \,|\, Z = z \sim \sigma^2 \text{Chi}^2(n-K).$$

We can obtain an unbiased estimator for σ^2 by dividing SSR by the degrees of freedom:

$$\hat{\sigma}^2 \equiv \frac{\text{SSR}}{n-K}, \tag{7}$$
$$E(\hat{\sigma}^2 \,|\, Z = z) = \sigma^2.$$

4. CONFIDENCE INTERVAL

We shall obtain a confidence interval for a linear combination of the coefficients:

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

Define the *standard error*

$$\text{SE} = [\hat{\sigma}^2 l'(x'x)^{-1} l]^{1/2}.$$

Our confidence interval will be based on the t-distribution.

Definition 2. If the random variables W and S are independent, with $W \sim \mathcal{N}(0,1)$, $S \sim \text{Chi}^2(m)$, then

$$\frac{W}{(S/m)^{1/2}} \sim t(m).$$

Claim 5.

$$\frac{l'(b-\beta)}{\text{SE}} \,|\, Z = z \;\sim\; t(n-K).$$

Proof. Conditional on $Z = z$,

$$\frac{l'(b-\beta)}{[\sigma^2 l'(x'x)^{-1}l]^{1/2}} \Bigg/ \left[\frac{\text{SSR}}{\sigma^2(n-K)}\right]^{1/2} \;\sim\; \mathcal{N}(0,1) \Bigg/ \left[\frac{\text{Chi}^2(n-K)}{n-K}\right]^{1/2}$$
$$\sim\; t(n-K),$$

where we have used the independence of b and SSR. We can cancel σ^2 in the numerator and denominator on the left-hand side, and simplify to obtain

$$\frac{l'(b-\beta)}{\text{SE}} \;\sim\; t(n-K). \quad \diamond$$

Note that because the conditional distribution of $l'(b-\beta)/\text{SE}$ given $Z = z$ does not depend on z, we have $l'(b-\beta)/\text{SE}$ independent of Z. The ratio $l'(b-\beta)/\text{SE}$ is called a *pivot* for $l'\beta$. It depends on the unknown parameters only through $l'\beta$, and it has a known distribution. This leads to a confidence interval for $l'\beta$.

The $t-$distribution is available in tables and in computer programs. Suppose that $n - K = 30$. We have

$$\text{Prob}(t(30) > 2.04) = .025,$$

and since the t-distribution is symmetric about zero,

$$\text{Prob}(|t(30)| \leq 2.04) = .95.$$

Then Claim 5 gives

$$\text{Prob}(-2.04 \leq \frac{l'\beta - l'b}{\text{SE}} \leq 2.04 \,|\, Z = z) = .95,$$

and so

$$\text{Prob}(l'b - 2.04 \cdot \text{SE} \leq l'\beta \leq l'b + 2.04 \cdot \text{SE} \,|\, Z = z) = .95. \tag{8}$$

Because the conditional probability in (8) does not depend on z, the unconditional probability is also equal to .95. We can write this as

$$\text{Prob}(\beta \in [l'b \pm 2.04 \cdot \text{SE}]) = .95.$$

As $n - K$ increases from 30 to infinity, the 97.5 percentile of the t-distribution decreases from 2.04 to a limiting value of 1.96 (which is the 97.5 percentile of a standard normal distribution).

5. CONFIDENCE ELLIPSE

We shall obtain a confidence region for two or more linear combinations of the coefficients. Let L be $h \times K$ so that $L\beta$ is $h \times 1$ and

$$Lb \,|\, Z = z \;\sim\; \mathcal{N}(L\beta, \sigma^2 L(x'x)^{-1}L').$$

Define

$$\hat{\text{Var}}(Lb) \equiv \hat{\sigma}^2 L(x'x)^{-1}L'.$$

Claim 6. If $W \sim \mathcal{N}(\mu, \Sigma)$ is $h \times 1$ and Σ is positive definite, then

$$(W - \mu)'\Sigma^{-1}(W - \mu) \;\sim\; \text{Chi}^2(h).$$

Proof. Since the $h \times h$ matrix Σ is positive definite and symmetric, there is a $h \times h$ matrix $\Sigma^{1/2}$ that is positive definite and symmetric such that

$$\Sigma = \Sigma^{1/2}\Sigma^{1/2}.$$

Then

$$\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2}$$

with $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

$$Q \equiv \Sigma^{-1/2}(W - \mu) \;\sim\; \mathcal{N}(0, \Sigma^{-1/2}\Sigma\Sigma^{-1/2}) = \mathcal{N}(0, I_h)$$

implies that

$$(W-\mu)'\Sigma^{-1}(W-\mu) = Q'Q = \sum_{j=1}^{h} Q_j^2 \sim \text{Chi}^2(h). \quad \diamond$$

Our confidence region for $L\beta$ will be based on the F-distribution.

Definition 3. If the random variables S_1 and S_2 are independent, with $S_1 \sim \text{Chi}^2(h)$ and $S_2 \sim \text{Chi}^2(m)$, then

$$\frac{S_1/h}{S_2/m} \sim F(h,m).$$

The parameters h and m of the F-distribution are called the numerator and denominator degrees of freedom.

Claim 7. Conditional on $Z = z$,

$$(Lb - L\beta)'[\hat{\text{Var}}(Lb)]^{-1}(Lb - L\beta)/h \sim F(h, n-K).$$

Proof. Conditional on $Z = z$,

$$\frac{(Lb - L\beta)'[\sigma^2 L(x'x)^{-1}L']^{-1}(Lb - L\beta)/h}{\text{SSR}/[\sigma^2(n-K)]} \sim \frac{\text{Chi}^2(h)/h}{\text{Chi}^2(n-K)/(n-K)} \sim F(h, n-K),$$

where we have used the independence of b and SSR. We can cancel σ^2 in the numerator and denominator on the left-hand side, and then simplify to obtain the result. $\diamond$

Note that because the conditional distribution of

$$(Lb - L\beta)'[\hat{\text{Var}}(Lb)]^{-1}(Lb - L\beta)/h \tag{9}$$

given $Z = z$ does not depend on z, the expression in (9) is independent of Z. The expression in (9) is a pivot for $L\beta$. It depends on the unknown parameters (β, σ^2) only through $L\beta$, and it has a known distribution. This leads to a confidence region for $L\beta$.

For example, suppose that $h = 2$ with

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} \quad \text{and} \quad L\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Suppose that $n-K=30$. The F-distribution is available in tables and computer programs. We have

$$\text{Prob}(F(2,30) > 3.32) = .05.$$

Then Claim 7 gives

$$\text{Prob}([\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} - \begin{pmatrix}b_1\\ b_2\end{pmatrix}]'[\hat{\text{Var}}\begin{pmatrix}b_1\\ b_2\end{pmatrix}]^{-1}[\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} - \begin{pmatrix}b_1\\ b_2\end{pmatrix}]/2 \leq 3.32) = .95. \qquad (10)$$

(The probability conditional on $Z=z$ is also .95.) The confidence region consists of the values for (β_1, β_2) that satisfy the inequality in (10). This gives the interior of an ellipse which is centered at the least-squares values (b_1, b_2).

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

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

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:

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

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

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

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

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

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 $\text{Cov}(G_i) = \Sigma$, 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 $E(G_i) = 0$, $\text{Cov}(G_i) = E(G_i G_i') = \Sigma$. 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,

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

Then Claim 3 gives

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

and so

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

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

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

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

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

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

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

The corresponding standard error is

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

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

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

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.

LECTURE NOTE 9

SYSTEM ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

1. INTRODUCTION

We shall work with the following general framework:

$$Q_i = R_i\gamma + V_i, \tag{1}$$

$$E(W_iV_i) = 0, \tag{2}$$

where we observe (Q_i, R_i, W_i) for $i = 1, \ldots, n$. As in Note 6, we shall assume random sampling, so that (Q_i, R_i, W_i) are independent and identically distributed (i.i.d.) from some joint distribution. The system aspect is that Q_i can be a vector: Q_i is $H \times 1$, R_i is $H \times K$, the parameter vector γ is $K \times 1$, and the vector of errors V_i is $H \times 1$. Estimation of γ is based on the orthogonality between W_i and the error V_i. The matrix W_i is $L \times H$, so W_iV_i is a $L \times 1$ vector, and $E(W_iV_i) = 0$ provides L orthogonality conditions to estimate the K components of γ. We need $L \geq K$. The variables in the matrix W_i are sometimes called "instrumental variables." In the current context, this just means that they generate orthogonality conditions.

A single linear predictor is a special case of this framework. Let $H = 1$ (not a system) and set $Q_i = Y_i$, $R_i = X_i'$, $W_i = X_i$. Then

$$Y_i = X_i'\gamma + V_i, \qquad E(X_iV_i) = 0$$

is equivalent to

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

The results obtained in this note will include the limit distribution for the least-squares estimator (Note 8) as a special case. In working with a single linear predictor, we saw

in Note 8 that we can do inference (in large samples) without making assumptions about the conditional variance of Y_i given X_i—we can allow for heteroskedasticity of a general, unknown form. That is still true in the system framework, but now there is an additional issue. The components of the error vector V_i may be correlated with each other, and our inference methods need to allow for that. The next section shows how our panel data models from Notes 4 and 5 can fit into the system framework. With longitudinal panel data, the correlation between components of V_i comes from correlation over time (serial correlation) in the errors for the same cross-section unit.

2. PANEL DATA

2.1 *Complete Conditioning*

First consider the case with complete conditioning, as in Note 4:

$$E(Y_{it} \mid Z_{i1}, \ldots, Z_{iT}, A_i) = Z'_{it}\gamma + A_i \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T). \tag{3}$$

For example, Y_{it} is the log of output for firm i in year t and Z_{it} is a $K \times 1$ vector containing measured inputs, such as capital and labor, for firm i in year t. The variable A_i is not observed; it is a firm effect that captures unmeasured inputs and differences in productivity that are constant over time, with variation only across the firms. This model for the conditional expectation imposes the exclusion restriction that only Z_{it} matters once we control for A_i—the past and future values of the measured inputs are excluded.

Define a prediction error:

$$U_{it} = Y_{it} - E(Y_{it} \mid Z_{i1}, \ldots, Z_{iT}, A_i), \tag{4}$$

so we can write the equations

$$Y_{it} = Z'_{it}\gamma + A_i + U_{it} \qquad (t = 1, \ldots, T).$$

The unmeasured variable A_i can create omitted variable bias. We can eliminate A_i by taking deviations from the time averages for each cross-section unit:

$$\bar{Y}_i \equiv \frac{1}{T}\sum_{t=1} Y_{it} = \bar{Z}'_i\gamma + A_i + \bar{U}_i,$$

$$Y_{it} - \bar{Y}_i = (Z_{it} - \bar{Z}_i)'\gamma + (U_{it} - \bar{U}_i) \quad (t = 1, \ldots, T). \tag{5}$$

The T deviation equations in (5) provide a system that maps into our framework in (1). Let

$$Q_i = \begin{pmatrix} Y_{i1} - \bar{Y}_i \\ \vdots \\ Y_{iT} - \bar{Y}_i \end{pmatrix}, \quad R_i = \begin{pmatrix} (Z_{i1} - \bar{Z}_i)' \\ \vdots \\ (Z_{iT} - \bar{Z}_i)' \end{pmatrix}, \quad V_i = \begin{pmatrix} U_{i1} - \bar{U}_i \\ \vdots \\ U_{iT} - \bar{U}_i \end{pmatrix}, \tag{6}$$

so that

$$Q_i = R_i\gamma + V_i.$$

Now we need orthogonality conditions. There is in fact considerable flexibility in setting up the W_i matrix. Here is one way to do it:

$$W_i = R_i' = (\,(Z_{i1} - \bar{Z}_i) \quad \ldots \quad (Z_{iT} - \bar{Z}_i)\,), \tag{7}$$

so that

$$W_iV_i = \sum_{t=1}^{T} (Z_{it} - \bar{Z}_i)(U_{it} - \bar{U}_i).$$

The orthogonality condition $E(W_iV_i) = 0$ follows from

$$E(Z_{it}U_{is}) = 0 \qquad (s, t = 1, \ldots, T).$$

Note that it would not be enough to just have Z_{it} orthogonal to U_{it} for each t, because we need U_{it} to be orthogonal to $\bar{Z}_i$. That U_{it} is orthogonal to $(Z_{i1}, \ldots, Z_{iT})$ follows from (3)—U_{it} is a prediction error, where we are conditioning on $Z_{i1}, \ldots, Z_{iT}$ (and on A_i). This is where complete conditioning is used. It would not be enough in (3) to condition only on Z_{it} and A_i. We need to condition on Z at all dates and then impose the exclusion restriction that only Z_{it} matters for Y_{it}. If there were no exclusion restrictions, we would have no leverage for dealing with the unobserved variable A_i.

Note that with this choice for W_i, we have $L = K$, so that the number of orthogonality conditions matches the number of elements in γ. It will turn out that our system estimator is simply the least-squares fit of $(Y_{it} - \bar{Y}_i)$ on $(Z_{it} - \bar{Z}_i)$, pooling all NT observations:

$$\hat{\gamma} = \arg\min_a \sum_{i=1}^{N} \sum_{t=1}^{T} [(Y_{it} - \bar{Y}_i) - (Z_{it} - \bar{Z}_i)'a]^2. \tag{8}$$

This pooled least-squares estimator using deviations has shown up before—see Section 5 of Note 4 on "Stacking."

There is a way to set up the W_i matrix that maximizes the number of orthogonality conditions. Let

$$Z_i = \begin{pmatrix} Z_{i1} \\ \vdots \\ Z_{iT} \end{pmatrix}$$

and set

$$W_i = \begin{pmatrix} Z_i & 0 & \dots & 0 \\ 0 & Z_i & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_i \end{pmatrix}. \tag{9}$$

Then

$$W_i V_i = \begin{pmatrix} Z_i(U_{i1} - \bar{U}_i) \\ \vdots \\ Z_i(U_{iT} - \bar{U}_i) \end{pmatrix} \tag{10}$$

and the number of orthogonality conditions is $L = T^2 \cdot K$. Now there are more orthogonality conditions than parameters to be estimated. Our estimator will allow for this by using a weight matrix.

Another way to eliminate A_i is to use first differences:

$$Y_{it} - Y_{i,t-1} = (Z_{it} - Z_{i,t-1})'\gamma + (U_{it} - U_{i,t-1}) \qquad (t = 2, \dots, T). \tag{11}$$

These $T - 1$ equations fit into the systems framework by setting

$$Q_i = \begin{pmatrix} Y_{i2} - Y_{i1} \\ \vdots \\ Y_{iT} - Y_{i,T-1} \end{pmatrix}, \quad R_i = \begin{pmatrix} (Z_{i2} - Z_{i1})' \\ \vdots \\ (Z_{iT} - Z_{i,T-1})' \end{pmatrix}, \quad V_i = \begin{pmatrix} U_{i2} - U_{i1} \\ \vdots \\ U_{iT} - U_{i,T-1} \end{pmatrix}. \tag{12}$$

For orthogonality conditions, we can set

$$W_i = R_i' = \left(\, (Z_{i2} - Z_{i1}) \quad \ldots \quad (Z_{iT} - Z_{i,T-1}) \, \right),$$

so that

$$W_i V_i = \sum_{t=2}^{T} (Z_{it} - Z_{i,t-1})(U_{it} - U_{i,t-1}).$$

This gives $L = K$, and our system estimator will turn out to be the least-squares fit of $(Y_{it} - Y_{i,t-1})$ on $(Z_{it} - Z_{i,t-1})$, pooling $N(T-1)$ observations:

$$\hat{\gamma} = \arg\min_a \sum_{i=1}^{N} \sum_{t=2}^{T} [(Y_{it} - Y_{i,t-1}) - (Z_{it} - Z_{i,t-1})'a]^2. \tag{13}$$

We can maximize the number of orthogonality conditions by using W_i as in (9) but with $T-1$ diagonal blocks, giving

$$W_i V_i = \begin{pmatrix} Z_i(U_{i2} - U_{i1}) \\ \vdots \\ Z_i(U_{iT} - U_{i,T-1}) \end{pmatrix} \tag{14}$$

and $T(T-1) \cdot K$ orthogonality conditions. This is less than the $T^2 \cdot K$ orthogonality conditions in (10), based on using deviations from time averages, but there is a linear dependence across the rows of $W_i V_i$ in (10):

$$Z_i(U_{i1} - \bar{U}_i) + \cdots + Z_i(U_{iT} - \bar{U}_i) = Z_i \sum_{t=1}^{T} (U_{it} - \bar{U}_i) = 0$$

since the deviations of U_{it} from the mean $\bar{U}_i$ sum to 0.

2.2 *Sequential Conditioning*

Consider the panel autoregression model from Note 5:

$$E(Y_{it} \,|\, Y_{i1}, \ldots, Y_{i,t-1}, A_i) = \lambda_t + \theta Y_{i,t-1} + A_i \qquad (t = 2, \ldots, T). \tag{15}$$

For example, Y_{it} is the log of earnings for individual i in year t, and we have earnings data for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The variable A_i is not observed; it is an individual effect that is constant over time with variation only across individuals. The model for Y_{it} conditions on all the past values back to Y_{i1} and assumes that only $Y_{i,t-1}$ matters, once we control for A_i.

Define the prediction error

$$U_{it} = Y_{it} - E(Y_{it} \,|\, Y_i^{(t-1)}, A_i), \tag{16}$$

with

$$Y_i^{(s)} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{is} \end{pmatrix}.$$

Then we can write the equations

$$Y_{it} = \lambda_t + \theta Y_{i,t-1} + A_i + U_{it} \qquad (t = 2, \ldots, T).$$

The sequential conditioning provides the following orthogonality conditions:

$$E(U_{it}) = 0, \quad E(Y_i^{t-1} U_{it}) = 0 \qquad (t = 2, \ldots, T). \tag{17}$$

We can eliminate A_i by taking first differences:

$$Y_{it} - Y_{i,t-1} = (\lambda_t - \lambda_{t-1}) + \theta(Y_{i,t-1} - Y_{i,t-2}) + (U_{it} - U_{i,t-1}) \qquad (t = 3, \ldots, T). \tag{18}$$

These $T - 2$ equations provide a system that maps into our framework in in (1). Let

$$Q_i = \begin{pmatrix} Y_{i3} - Y_{i2} \\ \vdots \\ Y_{iT} - Y_{i,T-1} \end{pmatrix}, \quad R_i = \begin{pmatrix} 1 & \ldots & 0 & Y_{i2} - Y_{i1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 1 & Y_{i,T-1} - Y_{i,T-2} \end{pmatrix}, \tag{19}$$

$$V_i = \begin{pmatrix} U_{i3} - U_{i2} \\ \vdots \\ U_{iT} - U_{i,T-1} \end{pmatrix}, \quad \gamma = \begin{pmatrix} \lambda_3 - \lambda_2 \\ \vdots \\ \lambda_T - \lambda_{T-1} \\ \theta \end{pmatrix}, \tag{20}$$

so that

$$Q_i = R_i\gamma + V_i.$$

For orthogonality conditions, we can use

$$W_i = \begin{pmatrix} I_{T-2} \\ W_{i2} \end{pmatrix}, \tag{21}$$

where I_{T-2} is the $(T-2)\times(T-2)$ identity matrix and

$$W_{i2} = \begin{pmatrix} Y_i^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Y_i^{(T-2)} \end{pmatrix}. \tag{22}$$

Then

$$W_iV_= \begin{pmatrix} U_{i3} - U_{i2} \\ \vdots \\ U_{iT} - U_{i,T-1} \\ Y_i^{(1)}(U_{i3} - U_{i2}) \\ \vdots \\ Y_i^{(T-2)}(U_{iT} - U_{i,T-1}) \end{pmatrix}. \tag{23}$$

Note that the orthogonality conditions in (17) imply that

$$E[Y_i^{(t-2)}(U_{it} - U_{i,t-1})] = 0 \qquad (t = 3, \dots, T), \tag{24}$$

and so we have

$$E(W_iV_i) = 0.$$

3. MOMENT EQUATION

The estimator is based on the orthogonality condition $E(W_iV_i) = 0$. Multiply both sides of equation (1) by W_i and then average both sides of the equation over the sample:

$$W_iQ_i = W_iR_i\gamma + W_iV_i \tag{25}$$

$$\frac{1}{n}\sum_{i=1}^{n} W_iQ_i = \left(\frac{1}{n}\sum_{i=1}^{n} W_iR_i\right)\gamma + \frac{1}{n}\sum_{i=1}^{n} W_iV_i. \tag{26}$$

Let $S_{WQ} = \sum_i W_i Q_i / n$ and write (26) as

$$S_{WQ} = S_{WR}\gamma + S_{WV}. \tag{26'}$$

This *moment equation* is the key to obtaining our estimator and its properties. The law of large numbers implies that

$$S_{WQ} - S_{WR}\gamma = S_{WV} \xrightarrow{p} E(W_i V_i) = 0,$$

which suggests the *minimum distance* estimator

$$\hat{\gamma} = \arg\min_a ||S_{WQ} - S_{WR}a||^2. \tag{27}$$

The norm in the distance criterion corresponds to the inner product

$$\langle a, b \rangle = a'\hat{C}b,$$

where $\hat{C}$ is a positive definite, symmetric matrix. The only requirement on the weight matrix $\hat{C}$ for our large sample results is that it converge in probability to a nonrandom matrix C, which is also positive definite and symmetric. As in Section 4 of Note 4, the solution to this minimum norm problem is a *generalized least-squares* estimator:

$$\hat{\gamma} = (S'_{WR}\hat{C}S_{WR})^{-1}S'_{WR}\hat{C}S_{WQ}. \tag{28}$$

Note that S_{WQ} on the left-hand side of (26′) is $L \times 1$, so that (26′) provides L equations to solve for the K elements in $\hat{\gamma}$, when we replace S_{WV} by its limit of 0. If $L = K$, then S_{WR} is a square matrix and, assuming it is nonsingular, we can solve for $\hat{\gamma}$:

$$\hat{\gamma} = S_{WR}^{-1}S_{WQ}. \tag{29}$$

In this case, (28) reduces to (29) for any $\hat{C}$:

$$\hat{\gamma} = S_{WR}^{-1}\hat{C}^{-1}(S'_{WR})^{-1}S'_{WR}\hat{C}S_{WQ} = S_{WR}^{-1}S_{WQ} \quad \text{if } L = K$$

(using $(AB)^{-1} = B^{-1}A^{-1}$ when A and B are $K \times K$ nonsingular matrices). When $L = K$, we are in the *just-identified* case.

So we only need a weight matrix in the *over-identified* case $L > K$. If we set

$$\hat{D} = S_{WR}\hat{C}, \tag{30}$$

then we can write (28) in a simpler form:

$$\hat{\gamma} = (\hat{D}S_{WR})^{-1}\hat{D}S_{WQ}.$$

This suggests another approach, which applies a weight matrix directly to the moment equation:

$$\hat{D}S_{WQ} = (\hat{D}S_{WR})\gamma + \hat{D}S_{WV}. \tag{31}$$

We require that $\hat{D}$ be $K \times L$, with $\hat{D}S_{WR}$ nonsingular, and that $\hat{D}$ converge in probability to a nonrandom matrix D, with $DE(W_iR_i)$ nonsingular. Then

$$\hat{D}S_{WQ} - (\hat{D}S_{WR})\gamma = \hat{D}S_{WV} \xrightarrow{p} D \cdot 0 = 0,$$

which suggests the following estimator:

$$\hat{\gamma} = (\hat{D}S_{WR})^{-1}\hat{D}S_{WQ}. \tag{32}$$

This coincides with the minimum-distance estimator if $\hat{D} = S'_{WR}\hat{C}$, but the weight matrix $\hat{D}$ need not have this form.

The next section shows that $\hat{\gamma}$ in (32) is a consistent estimator for γ, and Section 5 develops a limit distribution. These results using the $\hat{D}$ weight matrix will also apply to the minimum-distance form of the estimator in (28), using the $\hat{C}$ weight matrix.

4. CONSISTENT ESTIMATION

We shall work with the $\hat{D}$ form of $\hat{\gamma}$ in (32). Assume that $\hat{D} \xrightarrow{p} D$, a $K \times L$ nonrandom matrix, and that $DE(W_iR_i)$ is nonsingular. The proof is similar to the one for least squares in Note 8, Section 1.

Claim 1. $\hat{\gamma} \xrightarrow{p} \gamma$ as $n \to \infty$.

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

$$\hat{\gamma} \xrightarrow{p} [DE(W_i R_i)]^{-1} DE(W_i Q_i).$$

Multiply (25) by D and take expectations:

$$DW_i Q_i = DW_i R_i \gamma + DW_i V_i,$$
$$DE(W_i Q_i) = [DE(W_i R_i)]\gamma + 0.$$

Substitute into the probability limit for $\hat{\gamma}$ to obtain $\hat{\gamma} \xrightarrow{p} \gamma$. $\diamond$

5. LIMIT DISTRIBUTION

The limit distribution and the proof are similar to the results for least squares in Note 8, Section 2.

Claim 2. $\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(\alpha \Sigma \alpha')$ where

$$\alpha = [DE(W_i R_i)]^{-1} D, \quad \Sigma = E(W_i V_i V_i' W_i').$$

Proof. Substituting (26′) into (32) gives

$$\hat{\gamma} = (\hat{D} S_{WR})^{-1} \hat{D}(S_{WR}\gamma + S_{WV})$$
$$= \gamma + (\hat{D} S_{WR})^{-1} \hat{D} S_{WV},$$

so that

$$\sqrt{n}(\hat{\gamma} - \gamma) = (\hat{D} S_{WR})^{-1} \hat{D} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i V_i.$$

Define $G_i = W_i V_i$. Since G_i is a function of (Q_i, R_i, W_i), random sampling implies that the G_i are independent and identically distributed. In addition,

$$E(G_i) = 0, \quad \text{Cov}(G_i) = E(G_i G_i') = \Sigma,$$

and so the central limit theorem implies that

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

As in our proof that $\hat{\gamma}$ is a consistent estimator of γ, the law of large numbers and Slutsky (i) imply that

$$(\hat{D} S_{WR})^{-1} \hat{D} \xrightarrow{p} [DE(W_i R_i)]^{-1} D = \alpha.$$

Then Slutsky (ii) implies that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \alpha \cdot \mathcal{N}(0, \Sigma) = N(0, \alpha \Sigma \alpha'). \quad \diamond$$

6. CONFIDENCE INTERVAL

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

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

We are going to follow the argument for least squares in Note 8, Section 3 to use the limit distribution to construct a confidence interval. In order to obtain a consistent estimate of Λ, let $\hat{V}_i = Q_i - R_i \hat{\gamma}$ and define

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

with

$$\hat{\alpha} = (\hat{D} S_{WR})^{-1} \hat{D}, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} W_i \hat{V}_i \hat{V}_i' W_i'.$$

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

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

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

$$l' \gamma = \sum_{j=1}^{K} l_j \gamma_j.$$

Slutsky (i) implies that

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

and Slutsky (ii) implies that

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

Define the standard error as

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

We have established

Claim 3.

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

The ratio $l'(\hat{\gamma}-\gamma)/\text{SE}$ is an asymptotic pivot for $l'\gamma$. It depends upon the unknown parameters only through $l'\gamma$, and it has a known limit distribution. This leads to a confidence interval for $l'\gamma$. Claim 3 gives

$$\lim_{n\to\infty} \text{Prob}(-1.96 \leq \frac{l'\gamma - l'\hat{\gamma}}{\text{SE}} \leq 1.96) = .95,$$

and so

$$\lim_{n\to\infty} \text{Prob}(l'\hat{\gamma} - 1.96 \cdot \text{SE} \leq l'\gamma \leq l'\hat{\gamma} + 1.96 \cdot \text{SE}) = .95,$$

or

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

7. CONFIDENCE ELLIPSE

We can follow the procedure for the normal linear model in Note 7, Section 5. Let L be $h \times K$ so that $L\gamma$ is $h \times 1$. Then Slutsky (ii) implies that

$$L\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} L \cdot \mathcal{N}(0,\Lambda) = \mathcal{N}(0, L\Lambda L')$$

and

$$S_n \equiv (L\hat{\Lambda}L')^{-1/2} L\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, I_h). \tag{33}$$

Since S_n converges in distribution to $\mathcal{N}(0, I_h)$, the sum of squares $S'S$ converges in distribution to $\text{Chi}^2(h)$. This follows from another part of the Slutsky theorem, which is also known as the *continuous mapping theorem*:

Slutsky Theorem. (iii) Let S_n be a sequence of $h \times 1$ random variables that converges in distribution; let S be a random variable whose distribution is the limit distribution of S_n: $S_n \xrightarrow{d} S$. Then if $g : \mathcal{R}^h \to \mathcal{R}^m$ is a continuous function, $g(S_n) \xrightarrow{d} g(S)$.

Define

$$\hat{\text{Var}}(L\hat{\gamma}) = L\hat{\Lambda}L'/n.$$

Claim 4. $(L\hat{\gamma} - L\gamma)'[\hat{\text{Var}}(L\hat{\gamma})]^{-1}(L\hat{\gamma} - L\gamma) \xrightarrow{d} \text{Chi}^2(h)$.

Proof.

$$S_n'S_n = (L\hat{\gamma} - L\gamma)'(L\hat{\Lambda}L'/n)^{-1}(L\hat{\gamma} - L\gamma);$$

(33) and Slutsky (iii) imply that

$$S_n'S_n \xrightarrow{d} \text{Chi}^2(h). \quad \diamond$$

Claim 4 provides an asymptotic pivot for $L\gamma$, because

$$(L\hat{\gamma} - L\gamma)'[\hat{\text{Var}}(L\hat{\gamma})]^{-1}(L\hat{\gamma} - L\gamma) \tag{34}$$

depends upon the unknown parameters only through $L\gamma$, and it has a known distribution. This leads to a confidence region for $L\gamma$.

For example, suppose that $h = 2$ with

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} \quad \text{and} \quad L\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

The Chi^2 distribution is available in tables and computer programs. We have

$$\text{Prob}(\text{Chi}^2(2) > 5.99) = .05.$$

Then Claim 4 gives

$$\lim_{n\to\infty} \text{Prob}([\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} - \begin{pmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{pmatrix}]'[\hat{\text{Var}}\begin{pmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{pmatrix}]^{-1}[\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} - \begin{pmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{pmatrix}] \leq 5.99) = .95. \tag{35}$$

The confidence region consists of the values for (γ_1, γ_2) that satisfy the inequality in (35). This gives the interior of an ellipse which is centered at $(\hat{\gamma}_1, \hat{\gamma}_2)$.

8. SERIAL CORRELATION AND THE HOMOSKEDASTIC CASE

So far our basic assumptions have been random sampling and the the orthogonality condition $E(W_i V_i) = 0$. In particular, we have *not* assumed that the conditional expectation of V_i given W_i is zero, or that the conditional covariance matrix of V_i given W_i is constant. Now we are going to see how the asymptotic inference simplifies if we do make these additional assumptions. This will make it easier to see how the asymptotic inference deals with correlation across the components of the error vector V_i, as would be caused by serial correlation in longitudinal panel data.

So assume that

$$\text{(i) } E(V_i \,|\, W_i) = 0 \quad \text{and} \quad \text{(ii) } \text{Cov}(V_i \,|\, W_i) = E(V_i V_i' \,|\, W_i) = \Omega.$$

Here Ω is a constant $T \times T$ matrix. This implies that Σ depends only upon second moments instead of fourth moments:

$$\begin{aligned}\Sigma &= E[E(W_i V_i V_i' W_i' \,|\, W_i)] \\ &= E[W_i E(V_i V_i' \,|\, W_i) W_i'] \\ &= E(W_i \Omega W_i').\end{aligned}$$

Because $\Omega = E(V_i V_i')$, a consistent estimate of Ω is

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \hat{V}_i \hat{V}_i',$$

and a consistent estimate of Σ is

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} W_i \hat{\Omega} W_i'.$$

In the panel data application in Section 2.1,

$$V_i = \begin{pmatrix} U_{i1} - \bar{U}_i \\ \vdots \\ U_{iT} - \bar{U}_i \end{pmatrix},$$

and the (s,t) element of Ω is

$$\Omega_{st} = E(V_{is}V_{it}) = E[(U_{is} - \bar{U}_i)(U_{it} - \bar{U}_i)] \qquad (s,t = 1,\ldots,T).$$

Serial correlation in the errors U_{it} shows up in $E(U_{is}U_{it}) \neq 0$. Even if there is no such serial correlation in the U's, the transformed errors V_{it} will have serial correlation because U_{it} is correlated with $\bar{U}_i$. Our estimator for Ω deals with this by using the sample covariance of the estimated errors:

$$\hat{\Omega}_{st} = \frac{1}{n}\sum_{i=1}^{n} \hat{V}_{is}\hat{V}_{it}.$$

LECTURE NOTE 10

OPTIMAL WEIGHT MATRIX

1. INTRODUCTION

As in Note 9, we shall work with the following general framework:

$$Q_i = R_i\gamma + V_i, \tag{1}$$

$$E(W_iV_i) = 0, \tag{2}$$

where we observe (Q_i, R_i, W_i) for $i = 1, \ldots, n$. As in Note 6, we shall assume random sampling, so that (Q_i, R_i, W_i) are independent and identically distributed (i.i.d.) from some joint distribution. The system aspect is that Q_i can be a vector: Q_i is $H \times 1$, R_i is $H \times K$, the parameter vector γ is $K \times 1$, and the vector of errors V_i is $H \times 1$. Estimation of γ is based on the orthogonality between W_i and the error V_i. The matrix W_i is $L \times H$, so W_iV_i is a $L \times 1$ vector, and $E(W_iV_i) = 0$ provides L orthogonality conditions to estimate the K components of γ. We need $L \geq K$.

Note 9, equation (26′) has the following moment equation:

$$S_{WQ} = S_{WR}\gamma + S_{WV},$$

which leads to the minimum distance estimator:

$$\begin{aligned}\hat{\gamma} &= \arg\min_a (S_{WQ} - S_{WR}a)'\hat{C}(S_{WQ} - S_{WR}a) \\ &= (S'_{WR}\hat{C}S_{WR})^{-1}S'_{WR}\hat{C}S_{WQ}.\end{aligned}$$

Here $\hat{C}$ is a positive definite, symmetric matrix that converges in probability to a nonrandom matrix C, which is positive definite and symmetric. By random sampling, W_iV_i is

i.i.d., and so

$$\mathrm{Cov}(S_{WQ} - S_{WR}\gamma) = \mathrm{Cov}(S_{WV}) = \frac{1}{n}\mathrm{Cov}(W_i V_i) = \frac{1}{n}\Sigma.$$

Suppose that Σ is proportional to an identity matrix. Then symmetry suggests that we simply use Euclidean distance, so that $\hat{C} = C = I$. We shall show in Section 2 that $C = I$ is in fact optimal in this special case. So in this special case, we should obtain $\hat{\gamma}$ from a least-squares fit of S_{WQ} on S_{WR}.

Now let Σ be a general, positive definite, symmetric matrix. It has a square root:

$$\Sigma = \Sigma^{1/2}\Sigma^{1/2},$$

where $\Sigma^{1/2}$ is positive definite, symmetric. Define

$$\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$$

and use it to transform the moment equation:

$$\Sigma^{-1/2}S_{WQ} = (\Sigma^{-1/2}S_{WR})\gamma + \Sigma^{-1/2}S_{WV},$$

which we can write as

$$\tilde{S}_{WQ} = \tilde{S}_{WR}\gamma + \tilde{S}_{WV}.$$

Then

$$\mathrm{Cov}(\tilde{S}_{WV}) = \Sigma^{-1/2}(\frac{1}{n}\Sigma)\Sigma^{-1/2} = \frac{1}{n}I_L.$$

Now we are in the special case and should obtain $\hat{\gamma}$ from a least-squares fit of $\tilde{S}_{WQ}$ on $\tilde{S}_{WR}$:

$$\begin{aligned}
\hat{\gamma} &= (\tilde{S}'_{WR}\tilde{S}_{WR})^{-1}\tilde{S}'_{WR}\tilde{S}_{WQ} \\
&= (S'_{WR}\Sigma^{-1/2}\Sigma^{-1/2}S_{WR})^{-1}S'_{WR}\Sigma^{-1/2}\Sigma^{-1/2}S_{WQ} \\
&= (S'_{WR}\Sigma^{-1}S_{WR})^{-1}S'_{WR}\Sigma^{-1}S_{WQ}.
\end{aligned}$$

So the optimal choice for the weight matrix in the general case is $C = \Sigma^{-1}$, and $\hat{C}$ should be a consistent estimate of Σ^{-1}:

$$\hat{C} = \hat{\Sigma}^{-1} \xrightarrow{p} \Sigma^{-1} = C.$$

An alternative is to apply a weight matrix $\hat{D}$ directly to the moment equation, as in Note 9, equations (30), (31), and (32). The optimal choice is

$$\hat{D} = S'_{WR}\hat{\Sigma}^{-1} \xrightarrow{p} [E(W_i R_i)]'\Sigma^{-1},$$

with

$$\hat{\gamma} = (\hat{D}S_{WR})^{-1}\hat{D}S_{WQ}.$$

2. OPTIMAL WEIGHT MATRIX WHEN $\Sigma = I$

Note 9, Section 5 shows in Claim 2 that the limit distribution is

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, \Lambda)$$

with $\Lambda = \alpha\Sigma\alpha'$ and

$$\alpha = [DE(W_i R_i)]^{-1}D, \quad \Sigma = E(W_i V_i V_i' W_i').$$

Define $B = E(W_i R_i)$. When $\Sigma = I$, we have

$$\Lambda = (DB)^{-1}DD'(B'D')^{-1}.$$

If $D = B'$, this reduces to

$$\Lambda^* = (B'B)^{-1}.$$

So we need to show that for any choice of D such that DB is nonsingular, $\Lambda \geq \Lambda^*$. Here the inequality means that $\Lambda - \Lambda^*$ is a positive semidefinite matrix, so that $l'(\Lambda - \Lambda^*)l \geq 0$ for any $K \times 1$ vector l.

Claim 1. For any $K \times L$ matrix D such that DB is nonsingular,

$$(DB)^{-1}DD'(B'D')^{-1} \geq (B'B)^{-1}.$$

Proof. First note that

$$I_L - B(B'B)^{-1}B' \geq 0,$$

because for any $L \times 1$ vector a,

$$a'(I_L - B(B'B)^{-1}B')a = a'(a - \hat{a}) = (a - \hat{a})'(a - \hat{a}) \geq 0,$$

with

$$\hat{a} = B(B'B)^{-1}B'a.$$

(This says that the sum of squared residuals from a least-squares fit of a on B is non-negative, and we have used $(a - \hat{a})$ orthogonal to $\hat{a}$.) If an $L \times L$ matrix G is positive semidefinite, then so is HGH' for any $K \times L$ matrix H, because

$$l'(HGH')l = a'Ga \geq 0,$$

where the $L \times 1$ vector $a = H'l$. So with

$$H = (DB)^{-1}D$$

(so $HB = I_K$), we have

$$H(I_L - B(B'B)^{-1}B')H' = (DB)^{-1}DD'(B'D')^{-1} - (B'B)^{-1} \geq 0. \quad \diamond$$

LECTURE NOTE 11

GENERALIZED METHOD OF MOMENTS (GMM)

1. INTRODUCTION

Generalized method of moments is a framework that extends the orthogonality condition framework in Note 9. Let $Data_i$ denote the variables observed for the i^{th} cross-section unit and assume random sampling: $Data_i$ i.i.d. for $i = 1, \ldots, n$. We are given a moment function $\psi(\,\cdot\,,\,\cdot\,)$. We assume that there is a unique point γ in some parameter space that satisfies the *key condition*:

$$E[\psi(Data_i, \gamma)] = 0. \tag{1}$$

Here ψ is $L \times 1$, γ is $K \times 1$, and $L \geq K$. We want to make inferences on γ.

This framework contains the orthogonality condition framework in Note 9 as a special case. The moment function is

$$\psi(Data_i, a) = W_i(Q_i - R_i a).$$

In Note 9, we had

$$Q_i = R_i\gamma + V_i, \quad E(W_i V_i) = 0.$$

So

$$E[\psi(Data_i, \gamma)] = E[W_i(Q_i - R_i\gamma)] = E(W_i V_i) = 0,$$

and the key condition is satisfied. The special aspect of the moment function $\psi(Data_i, a) = W_i Q_i - W_i R_i a$ is that it depends upon the parameter a in a linear way. (For a given value of the first argument, $\psi(\,\cdot\,,\,\cdot\,)$ is an affine function of the second argument.) So GMM extends the framework to allow for moment functions that are nonlinear in the parameter. We shall need this extension in our discussion of likelihood methods in Note 13.

2. GMM ESTIMATOR

Define

$$g(a) = \frac{1}{n}\sum_{i=1}^{n} \psi(Data_i, a).$$

Then by the law of large numbers,

$$g(\gamma) \xrightarrow{p} E[\psi(Data_i, \gamma)] = 0 \quad \text{as} \quad n \to \infty.$$

This suggests obtaining an estimator $\hat{\gamma}$ from

$$\hat{\gamma} = \arg\min_{a} g(a)'\hat{C}g(a), \tag{2}$$

where $\hat{C}$ converges in probability to a nonrandom $L \times L$ matrix C, which is positive definite and symmetric. The first-order condition for the minimization in (2) is

$$(\partial g(\hat{\gamma})'/\partial a)\hat{C}g(\hat{\gamma}) = 0.$$

So the estimator satisfies

$$\hat{D}g(\hat{\gamma}) = 0, \tag{3}$$

with $\hat{D} = (\partial g(\hat{\gamma})'/\partial a)\hat{C}$. The condition in (3) will be very useful in obtaining the limit distribution of the estimator.

3. LIMIT DISTRIBUTION

Assume that $\hat{\gamma}$ is a consistent estimate of γ: $\hat{\gamma} \xrightarrow{p} \gamma$, and that it satisfies

$$\hat{D}g(\hat{\gamma}) = 0,$$

where $\hat{D}$ converges in probability to a $K \times L$ nonrandom matrix D, which satisfies the rank condition

$$DE[\frac{\partial \psi(Data_i, \gamma)}{\partial a'}] \quad \text{nonsingular.}$$

Then we have

$$0 = \hat{D}g(\hat{\gamma}) = \hat{D}[g(\gamma) + \frac{\partial g(\gamma^*)}{\partial a'}(\hat{\gamma} - \gamma)], \tag{4}$$

where, by the mean value theorem, this expansion holds for some point γ^* on the line segment connecting $\hat{\gamma}$ and γ. (There is a different point γ^* for each component of g.) By the central limit theorem,

$$\sqrt{n}g(\gamma) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = \text{Cov}(\psi(Data_i, \gamma)) = E[\psi(Data_i, \gamma)\psi(Data_i, \gamma)'].$$

From (4),

$$\sqrt{n}(\hat{\gamma} - \gamma) = -[\hat{D}\frac{\partial g(\gamma^*)}{\partial a'}]^{-1}\hat{D}\sqrt{n}g(\gamma).$$

Because γ^* is on the line segment connecting $\hat{\gamma}$ and γ, and $\hat{\gamma} \xrightarrow{p} \gamma$, we have $\gamma^* \xrightarrow{p} \gamma$. Because at any fixed value a, $(\partial g(a)/\partial a')$ converges in probability to $E[\partial\psi(Data_i, a)/\partial a']$, under regularity conditions we can obtain

$$[\hat{D}\frac{\partial g(\gamma^*)}{\partial a'}]^{-1}\hat{D} \xrightarrow{p} \left[DE[\frac{\partial\psi(Data_i, \gamma)}{\partial a'}]\right]^{-1} D \equiv \alpha.$$

Then Slutsky (ii) implies that

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} -\alpha \cdot \mathcal{N}(0, \Sigma) = \mathcal{N}(0, \alpha\Sigma\alpha').$$

We have provided a heuristic argument for the following

Claim 1. $\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, \Lambda)$ with $\Lambda = \alpha\Sigma\alpha'$ and

$$\alpha = \left[DE[\frac{\partial\psi(Data_i, \gamma)}{\partial a'}]\right]^{-1} D,$$
$$\Sigma = \text{Cov}(\psi(Data_i, \gamma)) = E[\psi(Data_i, \gamma)\psi(Data_i, \gamma)'].$$

4. OPTIMAL WEIGHT MATRIX

Following the argument in Note 10, it can be shown that the optimal choice for the weight matrix C is

$$C^* = \Sigma^{-1}.$$

The corresponding value for the weight matrix D is

$$D^* = E[\frac{\partial \psi(Data_i, \gamma)'}{\partial a}]\Sigma^{-1}.$$

With the optimal weight matrix, the asymptotic covariance matrix for $\hat{\gamma}$ is

$$\Lambda^* = \left[E[\frac{\partial \psi(Data_i, \gamma)'}{\partial a}]\Sigma^{-1}E[\frac{\partial \psi(Data_i, \gamma)}{\partial a'}]\right]^{-1}.$$

LECTURE NOTE 12

MINIMUM DISTANCE

1. INTRODUCTION

We first encountered a minimum-distance estimator in Section 4 of Note 4. Working with the panel data model with complete conditioning, we obtained a least squares fit corresponding to the linear predictor of Y_{it} given $Z_{i1}, \ldots, Z_{iT}$ and other variables. Doing this for $t = 1, \ldots, T$ resulted in a matrix $\hat{\Pi}$ of least-squares coefficients, which were then arranged in a vector $\hat{\pi}$. The model imposed restrictions on the population values π, but the least-squares estimates in $\hat{\pi}$ did not impose these restrictions. The restrictions were imposed using a minimum-distance estimator. The input for the minimum-distance estimator is not the original data on Y_{it} and Z_{it}, but rather the statistic $\hat{\pi}$ formed from the original data.

We also used a minimum-distance estimator in Section 4 of Note 5. Working with an autoregression model for panel data, we formed the sample covariances corresponding to the population covariances $\text{Cov}(Y_{is}, Y_{it})$ for $s, t = 1, \ldots, T$. These sample covariances were arranged in a vector $\hat{\sigma}$. The model imposed restrictions on the population values σ, but the sample covariances in $\hat{\sigma}$ did not impose these restrictions. The restrictions were imposed using a minimum-distance estimator. The input for the minimum-distance estimator is not the original data on Y_{it}, but rather the statistic $\hat{\sigma}$ formed from the original data.

This note sets up a general framework for minimum-distance estimation and provides a limit distribution that can be used for inference.

2. MINIMUM-DISTANCE ESTIMATOR

We are given a statistic $\hat{\pi}$ with a limit normal distribution centered at π:

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} \mathcal{N}(0, \Omega). \tag{1}$$

We are also given a distance function $h(\,\cdot\,,\,\cdot\,)$, which is continuously differentiable. We assume that there is a unique point γ in some parameter space that satisfies the *key condition*:

$$h(\pi, \gamma) = 0. \tag{2}$$

Here h is $L \times 1$, γ is $K \times 1$, and $L \geq K$. In Note 4, Section 4, the form of the distance function is $h(\hat{\pi}, a) = \hat{\pi} - G \cdot a$, where G is a given, known matrix (consisting of zeros and ones).

Because $h(\pi, \gamma) = 0$ and $\hat{\pi}$ is a consistent estimate of π, there is motivation for obtaining an estimator $\hat{\gamma}$ from

$$\hat{\gamma} = \arg\min_{a} h(\hat{\pi}, a)' \hat{C} h(\hat{\pi}, a), \tag{3}$$

where $\hat{C}$ converges in probability to a $L \times L$ nonrandom matrix C, which is positive definite and symmetric. The first-order condition for the minimization in (3) is

$$(\partial h(\hat{\pi}, \hat{\gamma})'/\partial a) \hat{C} h(\hat{\pi}, \hat{\gamma}) = 0.$$

So the estimator satisfies

$$\hat{D} h(\hat{\pi}, \hat{\gamma}) = 0, \tag{4}$$

with $\hat{D} = (\partial h(\hat{\pi}, \hat{\gamma})'/\partial a)\hat{C}$. The condition in (4) will be very useful in obtaining the limit distribution of the estimator.

3. LIMIT DISTRIBUTION

Assume that $\hat{\gamma}$ is a consistent estimate of γ: $\hat{\gamma} \xrightarrow{p} \gamma$, and that it satisfies

$$\hat{D} h(\hat{\pi}, \hat{\gamma}) = 0,$$

where $\hat{D}$ converges in probability to a $K \times L$ matrix D, which satisfies the rank condition

$$D\frac{\partial h(\pi,\gamma)}{\partial a'} \quad \text{nonsingular.}$$

The derivation of the limit distribution is similar to the argument used for GMM in Note 11. Apply the mean value theorem:

$$0 = \hat{D}h(\hat{\pi},\hat{\gamma}) = \hat{D}[h(\pi,\gamma) + \frac{\partial h(\pi^*,\gamma^*)}{\partial \pi'}(\hat{\pi}-\pi) + \frac{\partial h(\pi^*,\gamma^*)}{\partial a'}(\hat{\gamma}-\gamma),$$

where (π^*,γ^*) is on the line segment connecting $(\hat{\pi},\hat{\gamma})$ and (π,γ). Solving for $(\hat{\gamma}-\gamma)$ gives

$$\sqrt{n}(\hat{\gamma}-\gamma) = -[\hat{D}\frac{\partial h(\pi^*,\gamma^*)}{\partial a'}]^{-1}\hat{D}\frac{\partial h(\pi^*,\gamma^*)}{\partial \pi'}\sqrt{n}(\hat{\pi}-\pi).$$

Because $(\hat{\pi},\hat{\gamma}) \xrightarrow{p} (\pi,\gamma)$ and (π^*,γ^*) is on the line segment connecting $(\hat{\pi},\hat{\gamma})$ and (π,γ), we have $(\pi^*,\gamma^*) \xrightarrow{p} (\pi,\gamma)$. Then Slutsky (i) implies that

$$[\hat{D}\frac{\partial h(\pi^*,\gamma^*)}{\partial a'}]^{-1}\hat{D}\frac{\partial h(\pi^*,\gamma^*)}{\partial \pi'} \xrightarrow{p} [D\frac{\partial h(\pi,\gamma)}{\partial a'}]^{-1}D\frac{\partial h(\pi,\gamma)}{\partial \pi'}.$$

Define

$$\alpha = [D\frac{\partial h(\pi,\gamma)}{\partial a'}]^{-1}D.$$

Then Slutsky (ii) implies that

$$\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} -\alpha\frac{\partial h(\pi,\gamma)}{\partial \pi'}\mathcal{N}(0,\Omega) = \mathcal{N}(0,\alpha\Sigma\alpha'),$$

with

$$\Sigma = \frac{\partial h(\pi,\gamma)}{\partial \pi'}\Omega\frac{\partial h(\pi,\gamma)'}{\partial \pi}.$$

We have established

Claim 1. $\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} \mathcal{N}(0,\Lambda)$ with $\Lambda = \alpha\Sigma\alpha'$ and

$$\alpha = [D\frac{\partial h(\pi,\gamma)}{\partial a'}]^{-1}D,$$

$$\Sigma = \frac{\partial h(\pi,\gamma)}{\partial \pi'}\Omega\frac{\partial h(\pi,\gamma)'}{\partial \pi}.$$

4. OPTIMAL WEIGHT MATRIX

Following the argument in Note 10, it can be shown that the optimal weight matrix is

$$C^* = \Sigma^{-1}.$$

The corresponding value for the weight matrix D is

$$D^* = \frac{\partial h(\pi, \gamma)'}{\partial a} \Sigma^{-1}.$$

With the optimal weight matrix, the asymptotic covariance matrix for $\hat{\gamma}$ is

$$\Lambda^* = \Big[\frac{\partial h(\pi, \gamma)'}{\partial a} \Sigma^{-1} \frac{\partial h(\pi, \gamma)}{\partial a'} \Big]^{-1}.$$

5. DELTA METHOD

Suppose that $\gamma = g(\pi)$ for a given, known function g, which is continuously differentiable. As before we are given a statistic $\hat{\pi}$ with $\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} \mathcal{N}(0, \Omega)$. By Slutsky (i), we can obtain a consistent estimate of γ from $\hat{\gamma} = g(\hat{\pi})$. Then the minimum-distance framework can be used to obtain a limit distribution for $\hat{\gamma}$. The moment function is

$$h(\hat{\pi}, a) = g(\hat{\pi}) - a.$$

The minimum distance estimator is $\hat{\gamma} = g(\hat{\pi})$, so

$$h(\hat{\pi}, \hat{\gamma}) = 0$$

(and we can set $D = I$). Since $\alpha = -I$, Claim 1 gives

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

with

$$\Sigma = \frac{\partial g(\pi)}{\partial \pi'} \Omega \frac{\partial g(\pi)'}{\partial \pi}.$$

This is known as the *delta method*.

Ec. 2120 Spring 2010

G. Chamberlain

LECTURE NOTE 13

LIKELIHOOD

1. INTRODUCTION

As in Note 6, we shall assume random sampling:

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

This joint distribution implies a conditional distribution for Y_i conditional on $Z_i = z$. We specify a set of conditional distributions, indexed by a parameter θ, that contains this conditional distribution:

$$\text{Prob}(Y_i \in B \,|\, Z_i = z) = \int_B f(y \,|\, z, \theta)\, dm(y) \quad \text{for some} \quad \theta \in \Theta.$$

For each point θ in the parameter space Θ, there is a conditional density $f(\cdot \,|\, \cdot, \theta)$. The distribution of Y_i conditional on $Z_i = z$ has density $f(\cdot \,|\, z, \theta)$ for some θ in the parameter space. This density is with respect to the measure m. The function $f(\cdot \,|\, \cdot, \cdot)$ is given. It is known as the *likelihood function* (for a single observation).

For example, consider the normal linear model in Note 7:

$$Y_i \,|\, Z_i = z \;\sim\; \mathcal{N}(\beta' x, \sigma^2),$$

where $x = g(z)$ for a given, known function g. The parameter is $\theta = (\beta, \sigma^2)$, the parameter space is $\Theta = \mathcal{R}^K \times \mathcal{R}_+$, and the likelihood function is

$$f(y \,|\, z, (\beta, \sigma^2)) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp[-\frac{1}{2\sigma^2}(y - \beta' x)^2].$$

The model asserts that for some $\beta \in \mathcal{R}^K$ and $\sigma^2 \in \mathcal{R}_+$ (the true values),

$$\text{Prob}(Y_i \in [c, d] \mid Z_i = z) = \int_c^d f(y \mid z, (\beta, \sigma^2))\, dy.$$

(Here the measure m is Lebesgue measure on $\mathcal{R}$.) So we have a complete description of the possible conditional distributions for Y_i conditional on Z_i. (This is a partial likelihood, since we do not specify a set of distributions for Z_i; the marginal distribution of Z_i is left unrestricted.)

The normality assumption in the normal linear model is restrictive because there is a great variety of shapes for the density of a continuous distribution. Also the constant conditional variance assumption is restrictive. The conditional mean assumption that $E(Y_i \mid Z_i = z) = \beta' x$ need not be restrictive, because, for example, $\beta' x$ could represent a high-order polynomial.

If Y_i is a binary dependent variable, then a flexible specification for $E(Y_i \mid Z_i = z)$ implies a flexible specification for the conditional distribution of Y_i conditional on Z_i, and we can be more confident that our likelihood function is well specified. The next section considers a binary dependent variable.

2. BINARY DEPENDENT VARIABLE

2.1 *Probit Approximation*

Suppose that the dependent variable Y_i takes on only two values, which we shall denote by 0 and 1. Then the conditional expectation function is also a conditional probability function:

$$E(Y_i \mid Z_i) = 1 \cdot \text{Prob}(Y_i = 1 \mid Z_i) + 0 \cdot \text{Prob}(Y_i = 0 \mid Z_i) = \text{Prob}(Y_i = 1 \mid Z_i).$$

As in Note 2, let $r(\cdot)$ denote the regression function:

$$r(z) = E(Y_i \mid Z_i = z)$$

(which does not depend upon i due to random sampling). Since $r(\cdot)$ is also a conditional probability function, it only takes on values in the interval $[0, 1]$:

$$0 \leq r(\cdot) \leq 1.$$

As in Note 2, we can approximate the regression function by a linear predictor:

$$r(Z_i) \cong E^*(Y_i \,|\, X_{i1}, \ldots, X_{iK}) = \beta_1 X_{i1} + \ldots + \beta_K X_{iK} = \beta' X_i, \tag{1}$$

where X_{ik} is a given function of Z_i: $X_{ik} = g_k(Z_i)$. For example, we could use polynomial approximation; if Z_i is a scalar, we would have $X_{ik} = Z_i^{k-1}$. But since we know $r(\cdot)$ is between 0 and 1, we hope to get a better approximation (for a given number K of terms) by imposing this restriction.

We can impose the restriction by putting the linear approximation inside a given, known function that is bounded between 0 and 1. A popular choice is the probit function

$$\Phi(s) = \text{Prob}(W \leq s) \quad \text{where} \quad W \sim \mathcal{N}(0, 1).$$

So $\Phi(\cdot)$ is the cumulative distribution function (cdf) for the standard normal distribution. $\Phi(\cdot)$ is strictly monotonic with

$$\lim_{s \to -\infty} \Phi(s) = 0, \quad \lim_{s \to \infty} \Phi(s) = 1.$$

Now we can define a *probit approximation* to the regression function:

$$\gamma = \arg \min_{a \in \mathcal{R}^K} E[r(Z_i) - \Phi(a' X_i)]^2, \tag{2}$$

where the $K \times 1$ vector X_i is obtained from a given function of Z_i: $X_i = g(Z_i)$. As before, if Z_i is scalar, we could use $X_{ik} = Z_i^{k-1}$. Then our probit approximation is

$$r(Z_i) \cong \Phi(\gamma_1 X_{i1} + \ldots + \gamma_K X_{iK}) = \Phi(\gamma' X_i). \tag{3}$$

Its advantage over the linear predictor approximation in (1) is that it is guaranteed to stay between 0 and 1.

Define the prediction error

$$U_i = Y_i - E(Y_i \mid Z_i) = Y_i - r(Z_i),$$

and note that $E(U_i \mid Z_i) = 0$. Then

$$\begin{aligned} E[Y_i - \Phi(a'X_i)]^2 &= E[r(Z_i) + U_i - \Phi(a'X_i)]^2, \\ &= E[r(Z_i) - \Phi(a'X_i)]^2 + E(U_i^2), \end{aligned}$$

since

$$E[(r(Z_i)-\Phi(a'X_i))U_i] = E[E[(r(Z_i)-\Phi(a'X_i))U_i \mid Z_i]] = E[(r(Z_i)-\Phi(a'X_i))E(U_i \mid Z_i)] = 0.$$

So an equivalent definition of γ is

$$\gamma = \arg \min_{a \in \mathcal{R}^K} E[Y_i - \Phi(a'X_i)]^2. \tag{4}$$

The sample analog of the population definition of γ in (4) suggests the following estimator:

$$\hat{\gamma} = \arg \min_{a \in \mathcal{R}^K} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \Phi(a'X_i)]^2. \tag{5}$$

This is a *nonlinear least-squares* estimator.

2.2 *Partial Predictive Effect*

Once we have estimates for the probit approximation to the conditional expectation function, we can obtain predictive effects as in Section 4 of Note 2. For example, with two variables in Z_i, we have the following partial predictive effect from comparing the conditional expectation evaluated at $Z_{i1} = c$ and $Z_{i1} = d$, with Z_{i2} held constant at e:

$$E(Y_i \mid Z_{i1} = d, Z_{i2} = e) - E(Y_i \mid Z_{i1} = c, Z_{i2} = e) \cong \Phi(\gamma' g(d, e)) - \Phi(\gamma' g(c, e)), \tag{6}$$

with $X_i = g(Z_i)$. We obtain an estimate of this partial predictive effect of Z_1 on Y by replacing γ by the estimate $\hat{\gamma}$.

2.3 *Logit Approximation*

Another popular choice is the logit function

$$G(s) = \frac{\exp(s)}{1 + \exp(s)}.$$

Like the probit function, $G(\cdot)$ is strictly monotonic with

$$\lim_{s \to -\infty} G(s) = 0, \quad \lim_{s \to \infty} G(s) = 1.$$

$G(\cdot)$ is the cdf for the standard logistic distribution. This distribution is symmetric about 0 with the same general shape as a normal distribution, but its variance does not equal 1. We can define a *logit approximation* by replacing Φ by G in (2), (3), and (4). The coefficient vector γ will be different, but using the logit γ with G replacing Φ in (6) gives an approximation to the predictive partial effect which is usually quite similar to the probit approximation. Whether the probit or logit approximation will be better (for a given choice of $X_i = g(Z_i)$) will vary from one data set to another. Usually it does not matter and it is rarely if ever an important issue to focus on.

2.4 *Heteroskedasticity*

When Y_i takes on only the values 0 and 1, the conditional variance is determined by the conditional expectation:

$$\text{Var}(Y_i \mid Z_i) = E(Y_i^2 \mid Z_i) - [E(Y_i \mid Z_i)]^2 = E(Y_i \mid Z_i) - [E(Y_i \mid Z_i)]^2,$$

since $Y_i^2 = Y_i$. So

$$\text{Var}(Y_i \mid Z_i) = r(Z_i)[1 - r(Z_i)].$$

This suggests an alternative estimator for γ. First get a preliminary estimate $\hat{\gamma}^{(1)}$ using nonlinear least squares in (5); then form

$$\hat{r}_i(Z_i) = \Phi(\hat{\gamma}^{(1)'} X_i);$$

then do *weighted* (nonlinear) least-squares, using the inverse of the estimated conditional variance as a weight:

$$\hat{\gamma} = \arg \min_{a \in \mathcal{R}^K} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \Phi(a' X_i])^2 / [\hat{r}(Z_i)(1 - \hat{r}(Z_i))]. \tag{7}$$

The intuition for (7) is that observations Y_i with higher variance (conditional on Z_i) are given less weight in the fitting criterion.

2.5 *Likelihood Function*

Assume that the probit approximation is exact: $\text{Prob}(Y_i = 1 \,|\, Z_i) = \Phi(\gamma' X_i)$. Then the likelihood function is

$$f(y \,|\, z, \gamma) = \Phi(\gamma' x)^y [1 - \Phi(\gamma' x)]^{1-y} \quad \text{if} \quad y \in \{0, 1\},$$

with $x = g(z)$. If $y \notin \{0, 1\}$, then $f(y \,|\, z, \gamma) = 0$. The parameter space is $\Theta = \mathcal{R}^K$. The model asserts that for some $\gamma \in \mathcal{R}^K$ (the true value),

$$\text{Prob}(Y_i \in B \,|\, Z_i = z) = \sum_{y \in B} f(y \,|\, z, \gamma),$$

where B is a subset of $\{0, 1\}$. (Here the measure m is counting measure.)

3. INFORMATION INEQUALITY

To simplify notation, let (Y, Z) be a random vector with the F distribution: $(Y, Z) \sim F$, so that $(Y, Z) \stackrel{d}{=} (Y_i, Z_i)$. Let $E_\theta(\cdot \,|\, z)$ denote conditional expectation based on the $f(\cdot \,|\, z, \theta)$ density. In particular,

$$E_\theta\big(\log[f(Y \,|\, z, \tilde{\theta})] \,|\, z\big) = \int \log[f(y \,|\, z, \tilde{\theta})] f(y \,|\, z, \theta) \, dm(y).$$

Claim 1. (Information Inequality) For all $\theta, \tilde{\theta} \in \Theta$,

$$E_\theta\big(\log[f(Y \,|\, z, \tilde{\theta})] \,|\, z\big) \leq E_\theta\big(\log[f(Y \,|\, z, \theta)] \,|\, z\big).$$

Proof. Let Q denote the following random variable:

$$Q = f(Y \mid z, \tilde{\theta}) / f(Y \mid z, \theta).$$

By Jensen's inequality,

$$E_\theta\big(\log(Q) \mid z\big) \le \log\big(E_\theta(Q \mid z)\big).$$

Note that

$$\begin{aligned}
\log\big(E_\theta(Q \mid z)\big) &= \log \int \frac{f(y \mid z, \tilde{\theta})}{f(y \mid z, \theta)} f(y \mid z, \theta)\, dm(y) \\
&= \log \int f(y \mid z, \tilde{\theta})\, dm(y) \\
&= \log(1) = 0.
\end{aligned}$$

So

$$E_\theta\big(\log(Q) \mid z\big) = E_\theta\big(\log[f(Y \mid z, \tilde{\theta})] - \log[f(Y \mid z, \theta)] \mid z\big) \le 0,$$

which implies that

$$E_\theta\big(\log[f(Y \mid z, \tilde{\theta})] \mid z\big) \le E_\theta\big(\log[f(Y \mid z, \theta)] \mid z\big). \quad \diamond$$

4. MAXIMUM LIKELIHOOD IS CONSISTENT

The maximum-likelihood (ML) estimate of θ is

$$\hat{\theta} = \arg\max_{a \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i \mid Z_i, a). \tag{8}$$

By the information inequality, for any $a \in \Theta$,

$$\begin{aligned}
E\big(\log f(Y \mid Z, a)\big) &= E\Big(E_\theta\big(\log[f(Y \mid Z, a)] \mid Z\big)\Big) \\
&\le E\Big(E_\theta\big(\log[f(Y \mid Z, \theta)] \mid Z\big)\Big) \\
&= E\big(\log f(Y \mid Z, \theta)\big).
\end{aligned}$$

Under regularity conditions, we can obtain a *uniform law of large numbers*:

$$\sup_{a\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i \mid Z_i, a) - E\big(\log f(Y \mid Z, a)\big)| \xrightarrow{p} 0.$$

Then it can be shown that

$$\hat{\theta} = \arg\max_{a\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i \mid Z_i, a) \xrightarrow{p} \arg\max_{a\in\Theta} E\big(\log f(Y \mid Z, a)\big) = \theta.$$

5. LIMIT DISTRIBUTION FOR ML FROM GMM

The first-order condition for the maximization in (8) is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(Y_i \mid Z_i, \hat{\theta})}{\partial \theta} = 0.$$

Define

$$\psi((Y_i, Z_i), a) = \frac{\partial \log f(Y_i \mid Z_i, a)}{\partial \theta}.$$

This is known as the *score function*. We are going to use it as the moment function in GMM. Check the key condition:

$$\begin{aligned} E_\theta\Big(\frac{\partial \log f(Y_i \mid Z_i, \theta)}{\partial \theta} \mid Z_i = z\Big) &= \int [f(y \mid z, \theta)]^{-1} \frac{\partial f(y \mid z, \theta)}{\partial \theta} f(y \mid z, \theta)\, dm(y) \\ &= \frac{\partial}{\partial \theta} \int f(y \mid z, \theta)\, dm(y) \\ &= \frac{\partial}{\partial \theta} 1 = 0. \end{aligned}$$

So ψ is a a valid moment function.

Because $\dim(\psi) = \dim(\theta)$, we can set $\hat{D} = I$. So we have

Claim 1. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Lambda)$, with $\Lambda = \alpha\Sigma\alpha'$, and

$$\begin{aligned} \alpha &= \Big[E[\frac{\partial \psi((Y_i, Z_i), \theta)}{\partial \theta'}]\Big]^{-1}, \\ \Sigma &= E[\psi((Y_i, Z_i), \theta)\psi((Y_i, Z_i), \theta)']. \end{aligned}$$

6. INFORMATION EQUALITY

The argument used to show that the score function satisfies the key condition can be extended to show that

$$-E_\theta[\frac{\partial\psi((Y_i,Z_i),\theta)}{\partial\theta'} \mid Z_i = z] = E_\theta[\psi((Y_i,Z_i),\theta)\psi((Y_i,Z_i),\theta)' \mid Z_i = z].$$

This is known as the *information equality*. It implies that

$$-E[\frac{\partial\psi((Y_i,Z_i),\theta)}{\partial\theta'}] = E[\psi((Y_i,Z_i),\theta)\psi((Y_i,Z_i),\theta)'].$$

Hence $-\alpha = \Sigma^{-1}$. Combining this with Claim 1 gives

Claim 2. $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0,\Sigma^{-1})$.

7. PANEL PROBIT

7.1 *Latent Variable Crossing a Threshold*

The cross-section probit model can be expressed in terms of a latent variable Y_i^* crossing a threshold: $Y_i = 1(Y_i^* \geq 0)$, with

$$Y_i^* = X_i'\beta + U_i, \quad U_i \mid Z_i \sim \mathcal{N}(0,\sigma^2).$$

Then we have

$$\begin{aligned}\text{Prob}(Y_i = 1 \mid Z_i) &= \text{Prob}(U_i/\sigma \geq -X_i'(\beta/\sigma) \mid Z_i) \\ &= 1 - \Phi(-X_i'(\beta/\sigma)) \\ &= \Phi(X_i'\gamma),\end{aligned}$$

with $\gamma = \beta/\sigma$; β and σ are not separately identified.

7.2 *Random Effects*

Now suppose we have panel data:

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

with $Y_{it} = 0$ or 1. We assume random sampling for the cross-section units: (Y_i, Z_i) i.i.d. for $i = 1, \ldots, N$. The probit random-effects model can be obtained from a normal random-effects model for a latent variable Y_{it}^*.

$$\begin{aligned} Y_{it}^* &= X_{it}'\beta + U_{it}, \\ U_{it} &= V_i + \epsilon_{it} \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T), \end{aligned}$$

where, conditional on Z_i, V_i is independent of $(\epsilon_{i1}, \ldots, \epsilon_{iT})$ and

$$V_i \sim \mathcal{N}(0, \sigma_v^2), \quad \epsilon_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2) \quad (t = 1, \ldots, T).$$

(X_{it} is a given, known function of Z_i: $X_{it} = g_{it}(Z_i)$.)

Let $U_i' = (U_{i1}, \ldots, U_{iT})$. Then

$$U_i \,|\, Z_i \sim \mathcal{N}(0, \Omega)$$

with

$$\Omega = \sigma_v^2 1_T 1_T' + \sigma_\epsilon^2 I_T.$$

(1_T is a $T \times 1$ vector of ones.)

We observe $Y_{it} = 1$ if $Y_{it}^* \geq 0$; otherwise we observe $Y_{it} = 0$. Since only the sign of Y_{it}^* is observed, we can just as well work with $\tilde{Y}_{it}^* = Y_{it}^*/\sigma_\epsilon$:

$$\begin{aligned} \tilde{Y}_{it}^* &= X_{it}'\frac{\beta}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon}V_i + \frac{1}{\sigma\epsilon}\epsilon_{it} \\ &= X_{it}'\alpha + \tilde{V}_i + \tilde{\epsilon}_{it}, \end{aligned}$$

with

$$\alpha = \frac{\beta}{\sigma_\epsilon}, \quad \sigma_{\tilde{v}}^2 = \frac{\sigma_v^2}{\sigma_\epsilon^2}, \quad \sigma_{\tilde{\epsilon}}^2 = 1,$$

and

$$Y_{it} = \begin{cases} 1, & \text{if } \tilde{Y}_{it}^* \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

7.3 *Partial Effect*

In evaluating a partial effect, one possibility is to set $\tilde{V}_i = 0$, which is the mean, median, and mode of the $\tilde{V}_i$ distribution:

$$\text{Prob}(Y_{it} = 1 \,|\, X_{it} = s, \tilde{V}_i = 0) = \Phi(s'\alpha). \tag{9}$$

Another possibility is to average over the distribution of $\tilde{V}_i$:

$$E[\text{Prob}(Y_{it} = 1 \,|\, X_{it} = s, \tilde{V}_i)] = \int \Phi(s'\alpha + \tilde{v})h(\tilde{v})\, d\tilde{v}, \tag{10}$$

where h is the density for a $\mathcal{N}(0, \sigma^2_{\tilde{v}})$ distribution. Because $\tilde{V}_i$ is independent of X_{it}, we can use iterated expectations to evaluate (10):

$$\begin{aligned} E[E(Y_{it} \,|\, X_{it} = s, \tilde{V}_i)] &= E[E(Y_{it} \,|\, X_{it} = s, \tilde{V}_i) \,|\, X_{it} = s] \\ &= E(Y_{it} \,|\, X_{it} = s) \\ &= \text{Prob}((\tilde{V}_i + \tilde{\epsilon}_{it})/(\sigma^2_{\tilde{v}} + 1)^{1/2} \geq -s'\alpha/(\sigma^2_{\tilde{v}} + 1)^{1/2}) \\ &= \Phi(s'\alpha/(\sigma^2_{\tilde{v}} + 1)^{1/2}). \end{aligned} \tag{11}$$

Whether we use $\Phi(s'\alpha)$ from (9) or $\Phi(s'\alpha/(\sigma^2_{\tilde{v}} + 1)^{1/2})$ from (10) and (11) can make a big difference, because $\sigma_{\tilde{v}}$ can be arbitrarily large.

I think it is better to average over the distribution of $\tilde{V}_i$. When $\sigma_{\tilde{v}}$ is large, there is only a small fraction of the population with $\tilde{V}_i$ near 0.

Ec. 2120 Spring 2010

G. Chamberlain

LECTURE NOTE 14

INSTRUMENTAL VARIABLE MODEL

1. OMITTED VARIABLE BIAS

Suppose that we are interested in the long regression:

$$E(Y_i \mid FB_i, ED_i, A_i) = FB_i'\phi + ED_i\theta + A_i,$$

but data on A_i are not available, and we run a least-squares fit of Y on FB and ED. The least-squares coefficients will converge in probability to the coefficients in the following short linear predictor:

$$E^*(Y_i \mid FB_i, ED_i) = FB_i'\tilde{\phi} + ED_i\tilde{\theta}.$$

The relationship between the long coefficients ϕ, θ and the short coefficients $\tilde{\phi}$, $\tilde{\theta}$ is worked out in Note 3. We need to consider an auxiliary linear predictor of the omitted variable A_i on FB_i and ED_i:

$$E^*(A_i \mid FB_i, ED_i) = FB_i'\psi_1 + ED_i\psi_2.$$

The omitted variable formula gives

$$\tilde{\phi} = \phi + \psi_1, \quad \tilde{\theta} = \theta + \psi_2.$$

For example, Y_i is the log of earnings of individual i, FB_i consists of a constant and a set of family background variables, ED_i is years of schooling, and A_i is a measure of initial (prior to the schooling) ability. The scale of A_i is chosen so that its coefficient equals one in the long regression.

The short least-squares fit provides consistent estimates of the short linear predictor coefficients $\tilde{\phi}$ and $\tilde{\theta}$. But these differ from the long regression coefficients by the auxiliary coefficients ψ_1 and ψ_2. This is a classic problem of omitted variable bias. The instrumental variable model will provide a solution. This new model requires an additional variable (or set of variables) that satisfy certain exclusion restrictions.

2. EXCLUSION RESTRICTIONS AND RANDOM ASSIGNMENT

Now suppose that we observe an additional variable (or set of variables) SUB_i, so that we observe

$$(FB_i, SUB_i, ED_i, Y_i) \quad \text{for} \quad i = 1, \ldots, n.$$

A_i is not observed. As in Note 6, we assume random sampling. The first exclusion restriction is that SUB_i does not help to predict Y_i if it is added to the long regression:

$$E(Y_i \,|\, FB_i, SUB_i, ED_i, A_i) = FB_i'\phi + ED_i\theta + A_i.$$

The second exclusion restriction is that SUB_i does not help to predict A_i in a linear predictor that includes FB_i:

$$E^*(A_i \,|\, FB_i, SUB_i) = FB_i'\lambda.$$

For example, SUB_i is an education subsidy that provides encouragement to obtain additional schooling. So it is correlated with ED_i, but the first exclusion restriction is that once we control for ED_i (and the other regressors in the long regression), the amount of subsidy that the individual receives does not have any additional predictive power. The second exclusion restriction is satisfied if the subsidy is randomly assigned. Suppose that the subsidy takes on only two values, zero and one, and the value that is assigned to i is determined by a coin flip. Then SUB_i will not be correlated with A_i or FB_i, and so the partial correlation of A_i and SUB_i given FB_i will be zero. (See problem set 1 on partial correlation.)

Define the prediction errors

$$\epsilon_i = A_i - E^*(A_i \,|\, FB_i, SUB_i),$$
$$U_i = Y_i - E(Y_i \,|\, FB_i, SUB_i, ED_i, A_i),$$

and write the equations

$$A_i = FB_i'\lambda + \epsilon_i$$
$$Y_i = FB_i'\phi + Ed_i\theta + A_i + U_i.$$

Note that ϵ_i and U_i are orthogonal to FB_i and SUB_i. Substitute for A_i in the Y_i equation:

$$\begin{aligned} Y_i &= FB_i'(\phi + \lambda) + ED_i\theta + (\epsilon_i + U_i) \\ &= FB_i'\delta + ED_i\theta + V_i, \end{aligned}$$

with $\delta = \phi + \lambda$ and $V_i = \epsilon_i + U_i$. Note that FB_i and SUB_i are orthogonal to V_i:

$$E(FB_i \cdot V_i) = 0, \quad E(SUB_i \cdot V_i) = 0.$$

Now define

$$R_i = (\, FB_i' \quad ED_i \,), \quad W_i = \begin{pmatrix} FB_i \\ SUB_i \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}.$$

Then the exclusion restrictions imply that

$$Y_i = R_i\gamma + V_i, \quad E(W_iV_i) = 0. \tag{1}$$

This fits in the framework developed in Note 9. We can use results from Note 9 to obtain a consistent estimator for γ (provided that $E(W_iR_i)$ satisfies a rank condition). A consistent estimate of γ provides a consistent estimate of the coefficient θ on ED. The coefficient ϕ on FB in the long regression is not, however, consistently estimated. Instead we obtain a consistent estimate of $\delta = \phi + \lambda$. So there is still omitted variable bias in the FB coefficient (if FB_i and A_i are correlated).

The next section uses the orthogonality condition in (1) to obtain estimates and inferences that are valid in large samples.

3. JUST-IDENTIFIED CASE

Since ED_i is a scalar, the dimension K of the coefficient vector γ in (1) is

$$K = \dim(FB_i) + 1.$$

The dimension L of W_i is

$$L = \dim(FB_i) + \dim(SUB_i).$$

So if there is a single variable in SUB, then $L = K$ and the number of orthogonality conditions equals the number of parameters to be estimated. This is the *just-identified* case. The estimation of γ is based on the L orthogonality conditions in $E(W_i V_i) = 0$. The resulting estimator is often called an instrumental variables (IV) estimator. In the estimation context, all the variables in W are instrumental variables; there is no distinction between FB and SUB in providing orthogonality conditions. But in terms of the underlying model, FB and SUB play very different roles. The exclusion restrictions at the core of the model only apply to SUB. The random assignment argument only applies to SUB. So if we do refer to FB as instrumental variables (in the sense of generating orthogonality conditions), we should keep in mind that it is the *excluded* instrumental variables in SUB that play the key role in an instrumental variable model.

We can exploit the orthogonality conditions in (1) by multiplying the Y_i equation by W_i:

$$W_i Y_i = (W_i R_i)\gamma + W_i V_i,$$

$$E(W_i Y_i) = [E(W_i R_i)]\gamma,$$

and so

$$\gamma = [E(W_i R_i)]^{-1} E(W_i Y_i)$$

if $E(W_i R_i)$ is nonsingular. Then we can obtain a consistent estimate of γ by replacing population expectations by sample averages:

$$\begin{aligned}\hat{\gamma} &= \left(\frac{1}{n}\sum_{i=1}^{n} W_i R_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} W_i Y_i\right) \\ &= S_{WR}^{-1} S_{WY}.\end{aligned}$$

Suppose that $FB_i = 1$, so that (1) becomes

$$Y_i = \delta + ED_i\theta + V_i, \quad E(V_i) = 0 \quad E(SUB_i \cdot V_i) = 0.$$

Then $\text{Cov}(SUB_i, V_i) = 0$ and

$$\text{Cov}(SUB_i, Y_i) = \text{Cov}(SUB_i, ED_i)\theta.$$

We can solve for

$$\theta = \frac{\text{Cov}(SUB_i, Y_i)}{\text{Cov}(SUB_i, ED_i)}$$

if

$$\text{Cov}(SUB_i, ED_i) \neq 0.$$

So in addition to the exclusion restrictions on SUB, we require that SUB be correlated with ED. Then we can obtain a consistent estimate of θ by replacing the population covariances by their sample counterparts:

$$\hat{\theta} = \frac{\text{sample Cov}(SUB, Y)}{\text{sample Cov}(SUB, ED)}.$$

4. OVER-IDENTIFIED CASE

Now suppose there are two or more variables in SUB, so that $L > K$. This is the *over-identified* case. We still have

$$E(W_i Y_i) = E(W_i R_i)\gamma. \tag{2}$$

The rank condition on $E(W_iR_i)$ is that this $L \times K$ matrix has rank $= K$ (full column rank). This ensures that (2) determines γ uniquely. But in general we will not be able to solve for $\hat{\gamma}$ in the sample counterpart to (2), since $S_{WY} = S_{WR}\hat{\gamma}$ would give L equations for K unknowns. So we use a minimum-distance estimator:

$$\begin{aligned}\hat{\gamma} &= \arg\min_a (S_{WY} - S_{WR}a)'\hat{C}(S_{WY} - S_{WR}a) \\ &= (S'_{WR}\hat{C}S_{WR})^{-1}S'_{WR}\hat{C}S_{WY}.\end{aligned}$$

The only requirements on the $L \times L$ weight matrix $\hat{C}$ is that it be positive definite, symmetric and converge to a nonrandom matrix C that is positive definite, symmetric.

5. OPTIMAL WEIGHT MATRIX

From Note 10, the optimal choice for C is a matrix that is proportional to Σ^{-1}, where

$$\Sigma = \mathrm{Cov}(W_iV_i) = E(W_iV_iV_i'W_i') = E(V_i^2W_iW_i')$$

(since V_i is scalar). It is common to use a weight matrix that would be optimal under homoskedasticity. Then having chosen C, we use the general results in Note 9 for inference. So the standard errors, confidence sets, and p-values are valid in large samples without restricting the form of the heteroskedasticity.

The homoskedastic case has

$$\text{(i) } E(V_i \,|\, W_i) = 0,$$

$$\text{(ii) } \mathrm{Var}(V_i \,|\, W_i) = E(V_i^2 \,|\, W_i) = \sigma_v^2.$$

So the orthogonality condition $E(W_iV_i) = 0$ is strengthened to V_i mean-independent of W_i, and the conditional variance of V_i given W_i is assumed to be constant. Then

$$\Sigma = \sigma_v^2 E(W_iW_i').$$

Since we only need C to be proportional to Σ^{-1}, we can ignore σ_v^2 and use

$$\hat{C} = \left(\frac{1}{n}\sum_{i=1}^n W_iW_i'\right)^{-1} = S_{WW'}^{-1}.$$

Using this weight matrix gives

$$\hat{\gamma} = (S'_{WR}S^{-1}_{WW'}S_{WR})^{-1}S'_{WR}S^{-1}_{WW'}S_{WY}. \tag{3}$$

This is known as the two-stage least-squares estimator (TSLS or 2SLS). The two-stage interpretation comes from a different way of deriving the estimator, which is developed in the next section.

6. POPULATION: TWO-STAGE LINEAR PREDICTOR

Writing out the components of R_i in (1) gives

$$Y_i = FB'_i\delta + ED_i\theta + V_i, \quad E(W_iV_i) = 0. \tag{1'}$$

Use (1′) to form the linear predictor of Y_i given W_i:

$$E^*(Y_i\,|W_i) = FB'_i\delta + E^*(ED_i\,|\,W_i)\theta.$$

Define

$$ED^*_i = E^*(ED_i\,|\,W_i) = W'_i\tau.$$

Then the linear predictor of Y_i given FB_i and ED^*_i identifies δ and θ:

$$E^*(Y_i\,|\,FB_i, ED^*_i) = FB'_i\delta + ED^*_i\theta. \tag{4}$$

7. SAMPLE: TWO-STAGE LEAST SQUARES

From (4), a least-squares fit of Y on FB and ED^* would provide consistent estimates of δ and θ. The predicted value ED^*_i is orthogonal to the error V_i because ED^*_i is constructed from W_i, which is orthogonal to V_i. The TSLS estimator obtains a consistent estimate of τ in stage 1. This is a least-squares fit of ED on W, with fitted values $\widehat{ED}_i = W_i\hat{\tau}$. The second stage obtains consistent estimates of δ and θ from a least-squares fit of Y on FB and $\widehat{ED}$.

This two-stage least-squares estimator is in fact the same as the estimator in (3), based on an optimal weight matrix. In the first stage, we can form fitted values for each variable in R_i:

$$\hat{R}_i = \begin{pmatrix} FB_i' & \widehat{ED}_i \end{pmatrix} = W_i' S_{WW'}^{-1} S_{WR}.$$

We get a perfect fit for the FB variables, since they are included in W, but we can still use the formula for the least-squares fitted value. Then in the second stage, we have a least-squares fit of Y on $\hat{R}$:

$$\begin{aligned}
\hat{\gamma} &= \left(\frac{1}{n}\sum_{i=1}^{n} \hat{R}_i'\hat{R}_i)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{R}_i' Y_i \right. \\
&= [S_{WR}' S_{WW'}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i'\right) S_{WW'}^{-1} S_{WR}]^{-1} S_{WR}' S_{WW'}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i Y_i\right) \\
&= (S_{WR}' S_{WW'}^{-1} S_{WR})^{-1} S_{WR}' S_{WW'}^{-1} S_{WY}.
\end{aligned}$$

This is our orthogonality condition estimator with weight matrix $\hat{C} = S_{WW'}^{-1}$.

8. POTENTIAL OUTCOME FUNCTION, TREATMENT EFFECTS, SELECTION BIAS, AND RANDOM ASSIGNMENT

We shall use a potential outcome function to define an average treatment effect. Then we shall see how random assignment of the treatment allows us to obtain the average treatment effect from a predictive effect.

For each individual i, there is a *potential outcome function* $Y_i(\cdot)$. It can be evaluated at any feasible level t of the treatment. Then $Y_i(t)$ is a random variable, whose realized value is the outcome for i at treatment level t. As t varies, we have a set of potential outcomes. Only one of these potential outcomes is actually observed. Let T_i denote the treatment level that is assigned to i. Then the observed outcome is the potential outcome corresponding to the assigned treatment level:

$$Y_i = Y_i(T_i).$$

The *average treatment effect* in comparing treatment level t_1 with treatment level t_2 is

$$\text{ATE}(t_1, t_2) = E[Y_i(t_2) - Y_i(t_1)].$$

The corresponding predictive effect, as defined in Note 2, is

$$\begin{aligned}\text{PE}(t_1, t_2) &= E(Y_i \,|\, T_i = t_2) - E(Y_i \,|\, T_i = t_1) \\ &= E[Y_i(t_2) \,|\, T_i = t_2] - E[Y_i(t_1) \,|\, T_i = t_1].\end{aligned}$$

The predictive effect does not, in general, equal the average treatment effect, because the assigned treatment T_i may be correlated with potential outcomes. The difference between the predictive effect and the average treatment effect is called *selection bias*.

For example, suppose that the outcome is blood pressure and there are two treatments: $t = 0$ does nothing (a placebo) and $t = 1$ is a new drug. Each individual has two potential outcomes, $Y_i(0)$ and $Y_i(1)$. Suppose that the individuals who are assigned $T_i = 1$ have high blood pressure. Then $t = 1$ may lower blood pressure for each individual: $Y_i(1) - Y_i(0) < 0$, but $E[Y_i(1) \,|\, T_i = 1]$ is higher than $E[Y_i(0) \,|\, T_i = 0]$. Then the average treatment effect of the new drug is to lower blood pressure, but the predictive effect shows higher blood pressure on average for the treated $T_i = 1$ individuals compared with the untreated $T_i = 0$.

A solution to selection bias is *random assignment*. Suppose that each individual is assigned $T_i = 0$ or $T_i = 1$ based on a coin flip. Then T_i will be independent of the potential outcomes. Let $\mathcal{T}$ denote the set of possible values for the treatment. Our general definition of a randomly assigned treatment is that

$$\{Y_i(t),\ t \in \mathcal{T}\} \perp\!\!\!\perp T_i$$

—the random variables corresponding to the potential outcomes are jointly independent of the treatment assignment T_i. In the case of the placebo and the new drug, the treatment is randomly assigned if

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i.$$

Under random assignment, for any treatment levels t_1, t_2, and t:

$$E[Y_i(t_2) \,|\, T_i = t] = E[Y_i(t_2)], \quad E[Y_i(t_1) \,|\, T_i = t] = E[Y_i(t_1)],$$

and so the predictive effect equals the average treatment effect:

$$\begin{aligned} \mathrm{PE}(t_1, t_2) &= E(Y_i \,|\, T_i = t_2) - E(Y_i \,|\, T_i = t_1) \\ &= E[Y_i(t_2) \,|\, T_i = t_2] - E[Y_i(t_1) \,|\, T_i = t_1] \\ &= E[Y_i(t_2)] - E[Y_i(t_1)] \\ &= \mathrm{ATE}(t_1, t_2). \end{aligned}$$

The next section develops an instrumental variable model in which the treatment is not randomly assigned but there is an instrumental variable that is randomly assigned. We shall develop an orthogonality condition estimator for the average treatment effect, but this will require strong restrictions on the potential outcome function.

9. INSTRUMENTAL VARIABLE MODEL

In the drug example, suppose that individuals are randomly assigned to $t = 0$ (no treatment) and $t = 1$ (new drug), but the new drug has side effects and some of the people assigned to take it in fact do not take it. So there is a randomly assigned "intent to treat" but the actual treatment that i receives depends also on choices made by i. So there is the possibility of selection bias. The individuals who take the new drug in spite of the side effects may have potential outcomes that differ on average from the potential outcomes of the people who drop out.

More generally, suppose that T_i is not randomly assigned, but there is a variable (or set of variables) S_i that is randomly assigned and that is correlated with T_i. In the drug example, S could be the randomly assigned intent to treat. In the earnings and education example, S could be a randomly assigned education subsidy. We shall refer to S as a subsidy.

For each individual i, there is a potential outcome function $Y_i(\cdot\,,\cdot\,)$. It can be evaluated at any level t of the treatment and level s of the subsidy. Then $Y_i(t,s)$ is a random variable, whose realized value is the outcome for i at treatment level t and subsidy level s. As t and s vary, we have a set of potential outcomes. Only one of these potential outcomes is actually observed. Let T_i denote the treatment level assigned to i, and let S_i denote the subsidy level assigned to i. Then the observed outcome is the potential outcome corresponding to the assigned treatment and subsidy:

$$Y_i = Y_i(T_i, S_i).$$

A key exclusion restriction is that the distribution of the potential outcome $Y_i(t,s)$ does not depend on s: for all (feasible) values of t, s_1, and s_2,

$$Y_i(t,s_1) \stackrel{d}{=} Y_i(t,s_2).$$

(Here $\stackrel{d}{=}$ means that two random variables have the same distribution.) So we can write the potential outcome function as a function just of the treatment level:

$$Y_i(t,s) = Y_i(t), \quad Y_i = Y_i(T_i).$$

The definition of the subsidy being randomly assigned mimics the definition in Section 8 of a randomly assigned treatment. The random variables corresponding to the potential outcomes are jointly independent of the subsidy assignment:

$$\{Y_i(t),\ t \in \mathcal{T}\} \perp\!\!\!\perp S_i.$$

So there are two key assumptions in the instrumental variable model: the instrumental variable (or variables) S is excluded from the potential outcome function and S is randomly assigned. In order for our orthogonality condition estimator (also known as an IV estimator) to provide a consistent estimate of the average treatment effect, we need, in

addition to the two key assumptions, to restrict the form of the potential outcome function. Here is the restricted potential outcome function:

$$\begin{aligned} Y_i(t) &= Y_i(t_0) + \theta(t - t_0) \\ &= [Y_i(t_0) - \theta t_0] + \theta t \\ &= Y_{i0} + \theta t, \end{aligned}$$

where t_0 is some feasible treatment level and $Y_{i0} = Y_i(t_0) - \theta t_0$. So there is a linear response to the treatment level and the slope θ of the response does not vary across the individuals. The average treatment effect is

$$\text{ATE}(t_1, t_2) = \theta(t_2 - t_1),$$

which is the same as the treatment effect for each i. The heterogeneity across individuals is confined to the random intercept Y_{i0}.

Since S_i is randomly assigned,

$$E^*(Y_{i0} \mid 1, S_i) = E(Y_{i0}) \equiv \delta.$$

Define the prediction error

$$V_i = Y_{i0} - E^*(Y_{i0} \mid 1, S_i),$$

and note that V_i is orthogonal to 1 and to S_i. Then we have

$$Y_i(t) = \delta + \theta t + V_i,$$

and the observed outcome satisfies

$$Y_i = Y_i(T_i) = \delta + \theta T_i + V_i.$$

Just as with equation (1) in Section 2, we can put this in the form of the framework developed in Note 9:

$$Y_i = R_i\gamma + V_i, \quad E(W_i V_i) = 0,$$

with

$$R_i = (\,1 \quad T_i\,), \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}, \quad W_i = \begin{pmatrix} 1 \\ S_i \end{pmatrix}.$$

So we can use the orthogonality condition (IV) estimators developed in Note 9.

Even if S_i is not randomly assigned, we may be able to argue that, conditional on a set of variables Z_i, S_i is "as good as" randomly assigned. Given the restricted form of the potential outcome function, the assumption we need is that

$$E^*(Y_{i0} \mid Z_i, S_i) = Z_i'\delta,$$

so that S_i does not help to predict Y_{i0} in a linear predictor that includes Z_i. (Assume that Z_i includes a constant.) Define the prediction error

$$V_i = Y_{i0} - E^*(Y_{i0} \mid Z_i, S_i),$$

and note that V_i is orthogonal to Z_i and to S_i. Then we have

$$Y_i(t) = Z_i'\delta + \theta t + V_i,$$

and the observed outcome satisfies

$$Y_i = Y_i(T_i) = Z_i'\delta + \theta T_i + V_i.$$

So

$$Y_i = R_i\gamma + V_i, \quad E(W_iV_i) = 0, \tag{5}$$

with

$$R_i = (\,Z_i \quad T_i\,), \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}, \quad W_i = \begin{pmatrix} Z_i \\ S_i \end{pmatrix}.$$

Once again we can use the orthogonality condition (IV) estimators developed in Note 9.

Note the similarity of (5) with equation (1) in Section 2. Suppose that Y_i is the log of earnings for individual i, Z_i consists of a constant and a set of family background variables,

the treatment T_i is years of schooling, and S_i is an education subsidy. The selection bias arises because Y_{i0} contains A_i, a measure of initial ability that is not in the data set. V_i is the part of A_i that is not predictable from the family background variables. If V_i and T_i are correlated, then the coefficient on T_i in the linear predictor of Y_i given Z_i and T_i does not equal θ. Here the selection bias is equivalent to omitted variable bias.

10. REDUCED FORM

Another terminology for instrumental variables is *exogenous* variables. The variables Z and S in (5) are exogenous and the *endogenous* variables are the outcome Y and the assigned treatment T. The *reduced form* consists of either the conditional expectations or the linear predictors of the endogenous variables given the exogenous variables. In our IV model in (5), the linear predictors contain useful information:

$$E^*(T_i \mid Z_i, S_i) = Z_i'\alpha_1 + S_i'\pi_1, \tag{6}$$

$$\begin{aligned} E^*(Y_i \mid Z_i, S_i) &= Z_i'\delta + \theta(Z_i'\alpha_1 + S_i'\pi_1) \\ &= Z_i'(\delta + \theta\alpha_1) + S_i'(\theta\pi_1) \\ &= Z_i'\alpha_2 + S_i'\pi_2, \end{aligned} \tag{7}$$

where $\alpha_2 = \delta + \theta\alpha_1$ and

$$\pi_2 = \theta\pi_1. \tag{8}$$

The coefficients π_2 on S in predicting Y are proportional to the coefficients π_1 on S in predicting T, and the proportionality factor identifies θ.

Let $J = \dim(S_i)$. If $J = 2$,

$$\begin{pmatrix} \pi_{21} \\ \pi_{22} \end{pmatrix} = \theta \begin{pmatrix} \pi_{11} \\ \pi_{12} \end{pmatrix},$$

and the least-squares estimates of the linear predictors are

$$\hat{T}_i = Z_i'\hat{\alpha}_1 + S_{i1}\hat{\pi}_{11} + S_{i2}\hat{\pi}_{12},$$

$$\hat{Y}_i = Z_i'\hat{\alpha}_2 + S_{i1}\hat{\pi}_{21} + S_{i2}\hat{\pi}_{22}.$$

In this over-identified case, we can obtain two consistent estimates of θ from the least-squares estimates of the reduced form:

$$\hat{\theta}^{(1)} = \hat{\pi}_{21}/\hat{\pi}_{11}, \quad \hat{\theta}^{(2)} = \hat{\pi}_{22}/\hat{\pi}_{12}.$$

The two-stage least-squares estimator provides a way to combine consistent estimates in the over-identified case. There is an alternative minimum-distance estimator that directly imposes the key proportionality restriction in (8). Because of the proportionality restriction, we can express (π_1, π_2) as a function of a lower dimension, unrestricted parameter (θ, β):

$$\pi_1 = \beta, \quad \pi_2 = \theta\beta,$$

where π_1, π_2, and β are $J \times 1$ and θ is a scalar. The least-squares estimates $(\hat{\pi}_1, \hat{\pi}_2)$ will not satisfy the proportionality restriction in a finite sample, but the estimates are consistent and so converge in probability to (π_1, π_2), which do satisfy the restriction. The following minimum-distance estimator obtains consistent estimates of θ and β by imposing the proportionality restriction:

$$\begin{aligned}(\hat{\theta}, \hat{\beta}) &= \arg \min_{a \in \mathcal{R}, b \in \mathcal{R}^J} || \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - \begin{pmatrix} b \\ a \cdot b \end{pmatrix} ||^2 \\ &= \arg \min_{a \in \mathcal{R}, b \in \mathcal{R}^J} \left(\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - \begin{pmatrix} b \\ a \cdot b \end{pmatrix} \right)' \hat{C} \left(\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix} - \begin{pmatrix} b \\ a \cdot b \end{pmatrix} \right).\end{aligned}$$

Define

$$\hat{\pi} = \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix}, \quad \pi = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}.$$

The results from Note 9 can be used to show that

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

An optimal choice for C is Λ^{-1}:

$$\hat{C} = \hat{\Lambda}^{-1} \xrightarrow{p} \Lambda^{-1} = C.$$

The estimate $(\hat{\theta}, \hat{\beta})$ can be used to form estimates of the reduced-form coefficients that impose the proportionality restriction:

$$\hat{\pi}_1^* = \hat{\beta}, \quad \hat{\pi}_2^* = \hat{\theta} \cdot \hat{\beta}.$$

11. DEMAND FUNCTION

Let $D_i(p)$ denote the quantity demanded in market i at price p. The demand function $D_i(\cdot)$ can be evaluated at any price. Only one of these quantities is actually observed. Let P_i denote the observed price in market i. Then, assuming that the observed quantity Q_i is on the demand curve,

$$Q_i = D_i(P_i).$$

We shall work with a restricted form of the demand function:

$$\begin{aligned} D_i(p) &= D_i(p_0) + \theta(p - p_0) \\ &= [D_i(p_0) - \theta p_0] + \theta p \\ &= D_{i0} + \theta p, \end{aligned}$$

where p_0 is some reference price and $D_{i0} = D_i(p_0) - \theta p_0$. ($D_i(p)$ could be the log of the quantity demanded and p could be the log of price.) The goal is to estimate θ, the slope (or, in logs, the elasticity) of the demand curve. This slope is assumed to be the same in all markets. The heterogeneity across markets is confined to the intercept D_{i0}, which represents shifts in the demand curve.

Let $SUP_i(p)$ denote the quantity supplied in market i at price p, and suppose that the supply function has the following form:

$$SUP_i(p) = SUP_{i0} + \lambda p.$$

The heterogeneity across markets is confined to the intercept SUP_{i0}, which represents shifts in the supply curve.

Suppose that the observed price P_i is assigned to clear the market, equating the quantity demanded at P_i with the quantity supplied at P_i:

$$D_i(P_i) = SUP_i(P_i) = Q_i.$$

Then we can solve for

$$P_i = \frac{D_{i0} - SUP_{i0}}{\lambda - \theta}.$$

The predictive effect of price on quantity, comparing the prices p_1 and p_2, is

$$\begin{aligned} \mathrm{PE}(p_1, p_2) &= E(Q_i \mid P_i = p_2) - E(Q_i \mid P_i = p_1) \\ &= E(D_{i0} \mid P_i = p_2) - E(D_{i0} \mid P_i = p_1) + \theta(p_2 - p_1). \end{aligned}$$

This does not, in general, equal $\theta(p_2 - p_1)$ if the demand shift D_{i0} is correlated with the market clearing price P_i. Because

$$\mathrm{Cov}(D_{i0}, P_i) = \frac{\mathrm{Var}(D_{i0}) - \mathrm{Cov}(SUP_{i0}, D_{i0})}{\lambda - \theta},$$

there will, in general, be a correlation between the demand shift and the market clearing price. So the predictive effect does not correspond to the slope of the demand curve (or, in logs, the demand elasticity). This is a form of selection bias, since the price P_i is not randomly assigned. For any p, the assigned price P_i is correlated with $D_i(p)$ through its correlation with the demand shift D_{i0}. (This bias is also called a simultaneity bias, because the the observed price P_i and quantity Q_i are simultaneously determined by the intersection of the demand and supply curves for market i.)

There is an instrumental variable solution to this bias problem. The key exclusion restriction is

$$E^*(D_{i0} \mid Z_i, S_i) = Z_i'\delta.$$

Here Z_i consists of observed demand shift variables (and a constant), and S_i consists of observed supply shift variables. The excluded instrumental variables S_i are assumed to be

"as good as randomly assigned," in that they do not help to predict the demand shift D_{i0} in a linear predictor that includes Z_i.

Define the prediction error

$$V_i = D_{i0} - E^*(D_{i0} \mid Z_i, S_i),$$

and note that V_i is orthogonal to Z_i and S_i. Then we have

$$D_i(p) = Z_i'\delta + \theta p + V_i,$$

and the observed quantity satisfies

$$Q_i = D_i(P_i) = Z_i'\delta + \theta P_i + V_i.$$

So

$$Q_i = R_i\gamma + V_i, \quad E(W_i V_i) = 0, \tag{9}$$

with

$$R_i = (\, Z_i \quad P_i \,), \quad \gamma = \begin{pmatrix} \delta \\ \theta \end{pmatrix}, \quad W_i = \begin{pmatrix} Z_i \\ S_i \end{pmatrix}.$$

As with equation (5) in Section 9, we can use the orthogonality condition (IV) estimators developed in Note 9.

LECTURE NOTE 15

TREATMENT EFFECT HETEROGENEITY

1. POTENTIAL OUTCOMES AND SELECTION BIAS

As in Note 14, Section 9, there is a potential outcome function $Y_i(\cdot\,,\cdot\,)$. It can be evaluated at any level t of the treatment and level s of the subsidy. Then $Y_i(t,s)$ is a random variable, whose realized value is the outcome for i at treatment level t and subsidy level s. As t and s vary, we have a set of potential outcomes. Assume that the distribution of the potential outcome $Y_i(t,s)$ does not depend on s: for all (feasible) values of t, s_1, and s_2,

$$Y_i(t,s_1) \stackrel{d}{=} Y_i(t,s_2).$$

So we can write the potential outcome function as a function just of the treatment level:

$$Y_i(t,s) = Y_i(t).$$

Suppose that the treatment level takes on only two values, which we shall denote by $t=0$ and $t=1$. Let β_0 and β_1 denote the expected values of the two potential outcomes:

$$E[Y_i(0)] = \beta_0,$$

$$E[Y_i(1)] = \beta_1.$$

Define the prediction errors

$$U_{i0} = Y_i(0) - E[Y_i(0)],$$

$$U_{i1} = Y_i(1) - E[Y_i(1)],$$

giving two equations for the potential outcomes:

$$Y_i(0) = \beta_0 + U_{i0}, \tag{1}$$

$$Y_i(1) = \beta_1 + U_{i1}. \tag{2}$$

Let θ denote the average treatment effect:

$$\text{ATE} = \beta_1 - \beta_0 \equiv \theta.$$

Selection bias arises if the assigned treatment, T_i, is correlated with the potential outcomes. We need a flexible way to model this relationship and to allow T_i to be related to the assigned subsidy, S_i. We shall use a threshold crossing model, which compares a latent variable, V_i, to a function of the subsidy, $g(S_i)$:

$$T_i = \begin{cases} 1, & \text{if } V_i \leq g(S_i); \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Suppose that S_i consists of J variables: $S_i' = (S_{i1}, \ldots, S_{iJ})$. The function g maps a subset of $\mathcal{R}^J$ to the interval $[0, 1]$. The function g is not given; it is unknown and unrestricted except for being differentiable. Since g is unrestricted, it is not restrictive to assume that

$$V_i \,|\, S_i = s \sim \text{Uniform}[0, 1]. \tag{4}$$

Then we have

$$\text{Prob}(T_i = 1 \,|\, S_i = s) = \text{Prob}(V_i \leq g(s)) = g(s), \tag{5}$$

which is unrestricted.

The observed outcome Y_i is the potential outcome function $Y_i(\cdot)$ evaluated at the assigned (observed) treatment T_i:

$$\begin{aligned} Y_i = Y_i(T_i) &= Y_i(0) + T_i[Y_i(1) - Y_i(0)] \\ &= \beta_0 + T_i\theta + [U_{i0} + T_i(U_{i1} - U_{i0})]. \end{aligned} \tag{6}$$

The conditional expectation of observed outcome conditional on observed treatment is

$$E(Y_i \,|\, T_i) = \beta_0 + T_i\theta + E(U_{i0} \,|\, T_i) + T_i E(U_{i1} - U_{i0} \,|\, T_i).$$

The predictive effect of the treatment is

$$\begin{aligned} \text{PE} &= E(Y_i \,|\, T_i = 1) - E(Y_i \,|\, T_i = 0) \\ &= \theta + E(U_{i1} \,|\, T_i = 1) - E(U_{i0} \,|\, T_i = 0). \end{aligned} \tag{7}$$

The selection bias problem is that this predictive effect does not, in general, equal the average treatment effect if T_i is correlated with U_{i0} or with U_{i1}.

2. RANDOM ASSIGNMENT

If the treatment is randomly assigned, then it is independent of the potential outcomes:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i.$$

In that case,

$$E(U_{i1} \,|\, T_i = 1) = E(U_{i1}) = 0, \quad E(U_{i0} \,|\, T_i = 0) = E(U_{i0}) = 0,$$

and the predictive effect in (7) equals the average treatment effect.

Now suppose that the treatment is not randomly assigned, but the subsidy S_i is randomly assigned:

$$\{Y_{i0}, Y_{i1}, V_i\} \perp\!\!\!\perp S_i. \tag{8}$$

The reduced-form conditional expectations of the endogenous T and Y given the exogenous S will play a key role in the analysis.

$$E(T_i \,|\, S_i = s) = \text{Prob}(T_i = 1 \,|\, S_i = s) = g(s). \tag{9}$$

Here is the key step in obtaining $E(Y_i \,|\, S_i = s)$:

$$\begin{aligned} E[T_i(U_{i1} - U_{i0}) \,|\, S_i = s] &= E[E[T_i(U_{i1} - U_{i0}) \,|\, S_i = s, V_i] \,|\, S_i = s] \\ &= E[1(V_i \le g(s))E(U_{i1} - U_{i0} \,|\, V_i) \,|\, S_i = s] \\ &= \int_0^1 1(v \le g(s))E(U_{i1} - U_{i0} \,|\, V_i = v)\, dv \\ &= \int_0^{g(s)} E(U_{i1} - U_{i0} \,|\, V_i = v)\, dv. \end{aligned} \tag{10}$$

We have used

$$E(U_{i1} - U_{i0} \,|\, S_i = s, V_i) = E(U_{i1} - U_{i0} \,|\, V_i). \tag{11}$$

Because S_i is independent of $\{U_{i0}, U_{i1}, V_i\}$, it is independent of $\{U_{i0}, U_{i1}\}$ conditional on V_i, which implies (11). Now we can obtain $E(Y_i \mid S_i = s)$ from (6), using (9), (10), and $E(U_{i0} \mid S_i = s) = 0$. This gives

Claim 1. (Reduced Form)

$$E(T_i \mid S_i = s) = g(s), \tag{12}$$

$$E(Y_i \mid S_i = s) = \beta_0 + g(s)\theta + \int_0^{g(s)} E(U_{i1} - U_{i0} \mid V_i = v)\, dv. \tag{13}$$

3. MARGINAL TREATMENT EFFECT

We need to think about that latent variable V_i that is used to model treatment assignment. An important special case of our model has the assigned treatment T_i independent of the gain from treatment $(Y_i(1) - Y_i(0))$, conditional on S_i. Then V_i is independent of $(U_{i1} - U_{i0})$, and the reduced form for Y_i reduces to

$$E(Y_i \mid S_i = s) = \beta_0 + g(s)\theta. \tag{14}$$

In this case, the reduced form analysis for the (simple) IV model in Note 14, Section 10 becomes relevant:

$$E^*(T_i \mid 1, S_i) = \alpha_1 + S_i'\pi_1, \tag{15}$$

$$\begin{aligned} E^*(Y_i \mid 1, S_i) &= \beta_0 + E^*[g(S_i) \mid 1, S_i]\theta \\ &= \beta_0 + (\alpha_1 + S_i'\pi_1)\theta \\ &= (\beta_0 + \theta\alpha_1) + S_i'(\theta\pi_1) \\ &= \alpha_2 + S_i'\pi_2, \end{aligned} \tag{16}$$

where $\alpha_2 = \beta_0 + \theta\alpha_1$ and

$$\pi_2 = \theta\pi_1. \tag{17}$$

We have used

$$E^*[E(T_i \mid S_i) \mid 1, S_i] = E^*(T_i \mid 1, S_i)$$

—see Claim 1 in Section 2 of Note 2.

Now consider the general case where $E(U_{i1} - U_{i0} \,|\, V_i) \neq 0$. The *marginal treatment effect* is a function defined by

$$\text{MTE}(v) = E[Y_i(1) - Y_i(0) \,|\, V_i = v] = \theta + E(U_{i1} - U_{i0} \,|\, V_i = v).$$

The marginal treatment effect evaluated at v is the average treatment effect for the subpopulation with $V_i = v$. This subpopulation will have their treatment assignment change from 0 to 1 for a small change in the subsidy that has $g(s)$ go from a bit below v to a bit above v. The average treatment effect can be obtained by integrating the marginal treatment effect:

$$\text{ATE} = E[E(Y_i(1) - Y_i(0) \,|\, V_i)] = \int_0^1 \text{MTE}(v)\, dv.$$

In the general case, the reduced form in Claim 1 is more complex than (15) and (16). But the key insight coming out of the simple reduced form is that the average treatment effect is identified by a ratio of reduced form slopes. So we should look at a ratio of partial derivatives.

Claim 2.

$$\frac{\frac{\partial E(Y_i \,|\, S_i = s)}{\partial s_j}}{\frac{\partial E(T_i \,|\, S_i = s)}{\partial s_j}} = \text{MTE}(g(s)) \qquad (j = 1, \ldots, J).$$

Proof.

$$\frac{\partial E(Y_i \,|\, S_i = s)}{\partial s_j} = \frac{\partial g(s)}{\partial s_j} \cdot \theta + E(U_i - U_{i0} \,|\, V_i = g(s)) \cdot \frac{\partial g(s)}{\partial s_j},$$

$$\frac{\partial E(T_i \,|\, S_i = s)}{\partial s_j} = \frac{\partial g(s)}{\partial s_j}.$$

The ratio of these partial derivatives is

$$\theta + E(U_{i1} - U_{i0} \,|\, V_i = g(s)) = \text{MTE}(g(s)). \quad \diamond$$

Suppose that S consists of a single variable that has continuous variation over some interval. In order to apply Claim 2, we first need to obtain flexible approximations to the conditional expectation functions. This could be done using least-squares estimates of a linear predictor based on a polynomial (or some other series expansion) in S. With the binary variable T, we might use a polynomial inside a probit or logit function, and use a maximum likelihood estimator. Then we can take derivatives at various values for S, form ratios, and examine how the estimated marginal treatment effect varies over the interval (or intervals) where S has continuous variation.

This is not feasible if, for example, S takes on only two values. The next section applies our framework to that case.

4. LOCAL AVERAGE TREATMENT EFFECT

Suppose that the subsidy takes on only two values, which we shall denote by $s = 0$ and $s = 1$. Assume that the instrumental variable is relevant: $E(T_i \mid S_i = 1) \neq E(T_i \mid S_i = 0)$, and label the subsidy values so that $g(0) < g(1)$. Now the analog of a ratio of derivatives is the ratio of differences:

$$\begin{aligned}\frac{E(Y_i \mid S_i = 1) - E(Y_i \mid S_i = 0)}{E(T_i \mid S_i = 1) - E(T_i \mid S_i = 0)} &= \theta + \frac{1}{g(1) - g(0)} \int_{g(0)}^{g(1)} E(U_{i1} - U_{i0} \mid V_i = v)\, dv \quad (18)\\ &= \theta + E[U_{i1} - U_{i0} \mid g(0) \leq V_i \leq g(1)]\\ &= E[Y_i(1) - Y_i(0) \mid g(0) \leq V_i \leq g(1)]\\ &= \text{LATE}.\end{aligned}$$

If an individual has $g(0) < V_i \leq g(1)$, then the individual has $T_i = 1$ if $S_i = 1$ and $T_i = 0$ otherwise. If $V_i \leq g(0)$, then the individual has $T_i = 1$ regardless of the subsidy value (always taker), and if $g(1) < V_i$, then $T_i = 0$ regardless of the subsidy value (never taker). So the difference ratio in (18) identifies an average treatment effect for the compliers; this treatment effect is known as the *local average treatment effect*. It is the average treatment effect for the subpopulation for whom the subsidy has an effect.

The sample analog is the ratio of differences of sample means:

$$\frac{(\bar{Y} \mid S = 1) - (\bar{Y} \mid S = 0)}{(\bar{T} \mid S = 1) - (\bar{T} \mid S = 0)} \xrightarrow{p} \text{LATE}.$$

This estimator is known as the *Wald estimator*. Because S_i consists of a single binary variable, the Wald estimator equals the ratio $(\hat{\pi}_2/\hat{\pi}_1)$ of least-squares slope coefficients from the least-squares fits $\hat{T}_i = \hat{\alpha}_1 + S_i\hat{\pi}_1$ and $\hat{Y}_i = \hat{\alpha}_2 + S_i\hat{\pi}_2$. In this just-identified case, this equals the orthogonality condition (IV) estimator (sample $\text{Cov}(S, Y)$/sample $\text{Cov}(S, T)$), which equals the two-stage least-squares estimator.

FINAL REVIEW PROBLEMS 1

1. Vera knows what the normal linear model is, but she does not know the theory that leads to an exact confidence interval based on the t-distribution. Instead, she tries to use Monte Carlo simulation to provide a confidence interval. With

$$Y = x\beta + \sigma V, \quad V \mid X = x \sim \mathcal{N}(0, I_n),$$

and

$$b = (x'x)^{-1}x'Y, \quad e = Y - xb, \quad \text{SSR} = e'e,$$

she argues that the distribution of

$$(b - \beta)/\sqrt{\text{SSR}},$$

conditional on $X = x$, does not depend upon β or σ (although it does depend upon x).

(a) Is Vera correct so far? Explain.

(b) Now Vera uses Matlab to generate J matrices $Y^{(j)}$, each of which is $n \times 1$ and contains independent draws from the standard normal (N(0,1)) distribution. She calculates

$$b^{(j)} = (x'x)^{-1}x'Y^{(j)}, \quad e^{(j)} = Y^{(j)} - xb^{(j)}, \quad \text{SSR}^{(j)} = {e^{(j)}}'e^{(j)} \qquad (j = 1, ..., J),$$

forms the absolute values

$$|l'b^{(j)}/\sqrt{\text{SSR}^{(j)}}| \qquad (j = 1, ..., J),$$

and sorts them to find the 95^{th} percentile (.95 quantile) of their empirical distribution; i.e., she finds a value, which she labels CRIT, such that $.95 \cdot J$ of the absolute values are below CRIT and $.05 \cdot J$ of them are above CRIT. Then she argues that

$$[(l'b - \text{CRIT} \cdot \sqrt{\text{SSR}}),\ (l'b + \text{CRIT} \cdot \sqrt{\text{SSR}})]$$

provides an approximation to an exact .95 confidence interval for $l'\beta$, with an approximation error that goes to zero as $J \to \infty$. Is Vera correct? Explain.

2. Provide the Matlab code for question 9 in Computer Note 3. (Check the answer against the Stata output in Note 3.)

3. (a) Computer Note 4 uses Monte Carlo simulation to approximate the probability that the robust confidence interval covers the true value. The simulation uses $J = 1000$ draws. Explain the sense in which the approximation error goes to zero as the number of draws increases.

(b) Explain how the central limit theorem can be used to assess the accuracy of the approximation when $J = 1000$.

4. Computer Note 4 uses Monte Carlo simulation to evaluate the finite sample properties of the robust confidence interval (from Note 8) when the population generating the data in fact satisfies the normal linear model. Now consider a finite sample evaluation when there is heteroskedasticity. Your data are in the $n \times 1$ matrix y and the $n \times K$ matrix x. One possibility is to start by calculating

$$b = (x'x)^{-1}x'y, \quad e = y - xb,$$

and using a nonlinear least-squares program to obtain

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{i=1}^{n} [e_i^2 - \exp(x_i'\gamma)]^2.$$

Explain how to complete the analysis; provide motivation for your procedure.

FINAL REVIEW PROBLEMS 2

1. This problem extends the linear predictor concept to allow Y_i to be a vector. In addition, the dimension of Y_i may depend upon i. We shall develop inference that corresponds to the cluster option in Stata. Allowing the dimension of Y_i to depend upon i corresponds to allowing the number of observations within a cluster to vary across the clusters.

The key is random sampling of clusters:

$$(Y_i, X_i) \quad \text{i.i.d.} \qquad (i = 1, \ldots, n),$$

where Y_i is $H_i \times 1$ and X_i is $H_i \times K$. Here H_i, the size of cluster i, is a random variable. Due to the random sampling of clusters, H_i is i.i.d. Define the $K \times 1$ vector β to solve the following linear predictor problem:

$$\beta = \arg \min_{a \in \mathcal{R}^K} E[(Y_i - X_i a)'(Y_i - X_i a)].$$

(a) Define the prediction error

$$U_i = Y_i - X_i \beta.$$

Show that

$$E(X_i' U_i) = 0.$$

(b) Let $\hat{\beta}$ denote the least-squares estimator:

$$\hat{\beta} = (X'X)^{-1} X'Y,$$

where

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

The number of rows in X and Y is $\sum_{i=1}^n H_i$; X has K columns, Y has one. Show that

$$\hat{\beta} \xrightarrow{p} \beta \quad \text{as} \quad n \to \infty.$$

(c) Use the arguments in Note 9 to show that

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

Provide a formula for Λ.

(d) Explain how to use the limit distribution in (c) to obtain standard errors and confidence intervals corresponding to the ones Stata reports with the cluster option.

2. In class, we first developed the following special case of Note 9:

$$Q_i = R_i\gamma + V_i, \quad E(R_i'V_i) = 0 \qquad (i = 1, \ldots, n), \tag{1}$$

where Q_i is $H \times 1$ and R_i is $H \times K$. Then we developed the general case:

$$Q_i = R_i\gamma + V_i, \quad E(W_iV_i) = 0 \quad (i = 1, \ldots, n), \tag{2}$$

where W_i is $L \times H$ with $L \geq K$. Based on problem 1, it appears that we can regard the orthogonality condition $E(R_i'V_i)$ in (1) as holding by construction. But we used this framework to do inference with the Mundlak method, based on (within group) deviations. The justification was based on Note 4, where we developed the complete conditioning restrictions. If $E(R_i'V_i) = 0$ can hold by construction, it might seem that the complete conditioning restrictions are not needed. Are they needed? Explain.

3. It turns out that Vera's method (Final Review Problems 1, question 1) works fine. It is remarkable that all the benefits of the exact t-based confidence interval can be obtained simply by recognizing that

$$(b - \beta)/\sqrt{\text{SSR}}$$

is a pivot. Vera has been studying Note 8. She sees

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

and wonders whether she could construct a useful confidence interval just by treating

$$\frac{l'(b - \beta)}{\text{SE}}$$

as a pivot. When working with the normal linear model, she could form a population distribution by choosing values for β and σ. But in Note 8, the population distribution is left very general. So Vera tries using the empirical distribution of her sample data as a population. She forms a random sample of size n as follows. Draw a number at random from the set $\{1, \ldots, n\}$ (assigning probability $1/n$ to each of those integers); the data on that cross-section unit supplies the first observation. Take an independent draw from the same set $\{1, \ldots, n\}$; the data on that cross-section unit supplies the second observation. Repeat, sampling with replacement, to get a sample of size n. Use this sample to construct

$$\frac{l'(b^{(1)} - b)}{\text{SE}^{(1)}}.$$

She draws J independent samples in this way, forms the absolute values

$$\left|\frac{l'(b^{(j)} - b)}{\text{SE}^{(\text{j})}}\right| \qquad (j = 1, \ldots, J),$$

and sorts them to find the 95^{th} percentile (.95 quantile) of their empirical distribution; i.e., she finds a value, which she labels CRIT, such that $.95 \cdot J$ of the absolute values are below CRIT and $.05 \cdot J$ of them are above CRIT. Then she argues that

$$[(l'b - \text{CRIT} \cdot \text{SE}),\ (l'b + \text{CRIT} \cdot \text{SE})]$$

can be used as an approximate .95 confidence interval for $l'\beta$. Do you think this procedure can be justified, in the sense that as $J \to \infty$ and $n \to \infty$ the probability that the interval covers the true value converges to .95? Explain your reasoning.

FINAL REVIEW PROBLEMS 3

1. The purpose of this problem is to: (i) gain experience working with conditional probability by deriving Bayes's Theorem; (ii) develop a striking similarity between an approximation to Bayesian inference and the limit distribution for the maximum-likelihood estimator in Note 13; (iii) sharpen our interpretation of a frequentist confidence interval by comparing it to a conditional (posterior) probability interval; (iv) examine the relevance of the Note 13 theory for decision making by making a comparison with Bayesian (really Savage's) decision theory.

Let Y denote a random vector with distribution conditional on $Z = z$ equal to $\mathcal{P}_z$. (Below we shall specialize to the case of random sampling, with $Y = (Y_1, \ldots, Y_n)$, $Z = (Z_1, \ldots, Z_n)$, and (Y_i, Z_i) i.i.d. for $i = 1, \ldots, n$. For now, it is convenient to let (Y, Z) denote the complete observation, without insisting that it is the result of a random sample.) Suppose that we have specified a family of conditional distributions for Y conditional on $Z = z$ such that the (population) conditional distribution $\mathcal{P}_z$ is in this family. The family of conditional distributions is given by a family of conditional density functions, indexed by a parameter θ which takes on values in a parameter space Θ, which is a subset of $\mathcal{R}^K$. The densities $f(y \,|\, z, \theta)$ are with respect to a single measure m. The assumption that the family contains the population distribution means that there is some value $\theta^* \in \Theta$ such that

$$\Pr\{Y \in A \,|\, Z = z, \mathcal{P}_z\} = \mathcal{P}_z(A) = \int_A f(y \,|\, z, \theta^*)\, dm(y). \tag{1}$$

(In the continuous case, m is Lebesgue measure and we can replace $dm(y)$ by dy; in the discrete case, m is counting measure, and we can replace the integral by a sum. With random sampling, we would have

$$f(y \,|\, z, \theta) = \prod_{i=1}^{n} f(y_i \,|\, z_i, \theta),$$

and it might be wise to change the notation to $f^{(n)}(y \,|\, z, \theta)$, which is the likelihood function for the full sample, to avoid confusing it with the likelihood function for a single observation.)

Suppose that we introduce a distribution on the parameter space Θ. This is a personal, subjective distribution that represents uncertainty about the value of θ^*. This distribution is conditional on $Z = z$, but we shall consider the case where it does not depend on z:

$$\Pr\{\theta^* \in B \,|\, Z = z\} = \int_B \pi(\theta)\, d\theta.$$

This (prior) distribution has density π with respect to Lebesgue measure on $\mathcal{R}^K$.

Multiplying the conditional density $f(y \mid z, \theta)$ by the marginal density $\pi(\theta \mid z) = \pi(\theta)$ gives the joint density $f(y, \theta \mid z)$ (with all densities conditional on z):

$$\Pr\{Y \in A, \theta^* \in B \mid Z = z\} = \int_A \int_B f(y \mid z, \theta)\pi(\theta)\, dm(y)\, d\theta.$$

The joint density $f(y, \theta \mid z)$ can be factored the other way, into a product of the conditional density $\pi(\theta \mid z, y)$ and the marginal density $f(y \mid z)$ (with all densities still conditional on z):

$$\Pr\{Y \in A, \theta^* \in B \mid Z = z\} = \int_A \int_B \pi(\theta \mid z, y) f(y \mid z)\, dm(y)\, d\theta.$$

So we have

$$f(y, \theta \mid z) = f(y \mid z, \theta)\pi(\theta) = \pi(\theta \mid z, y) f(y \mid z),$$

which gives Bayes's Theorem:

$$\pi(\theta \mid z, y) = f(y \mid z, \theta)\pi(\theta)/f(y \mid z).$$

A convenient shorthand is

$$\pi(\theta \mid z, y) \propto f(y \mid z, \theta)\pi(\theta),$$

where it is understood that we can recover the normalizing constant because the density integrates to 1:

$$\int_\Theta \pi(\theta \mid z, y)\, d\theta = c \int f(y \mid z, \theta)\pi(\theta)\, d\theta = 1$$
$$\Rightarrow \quad c = 1 \Big/ \int f(y \mid z, \theta)\pi(\theta)\, d\theta = 1/f(y \mid z).$$

(c is a "constant" when we condition on the sample value (z, y).)

Let (z, y) denote the sample realized value of (Z, Y). Suppose that the log-likelihood function, regarded as a function of θ with (z, y) fixed at the sample value, is well approximated by the quadratic part of a Taylor series expansion around the maximum-likelihood estimate $\hat{\theta}$:

$$L(\theta) = \log f(y \mid z, \theta) \approx L(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})' H(\hat{\theta})(\theta - \hat{\theta}),$$

where $H(\theta)$ is the Hessian matrix for the log-likelihood function:

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}$$

(and the linear term $[\partial L(\hat{\theta})/\partial\theta'](\theta - \hat{\theta})$ drops out because of the first-order condition for the maximum-likelihood estimate $\hat{\theta}$). Suppose that the likelihood function $f(y \mid z, \theta)$ is quite concentrated around $\hat{\theta}$. Then we have

$$\pi(\theta) \approx \pi(\hat{\theta})$$

over the range of values for which the likelihood is not essentially equal to 0. So outside of a set B the likelihood is essentially 0, and the set B is small enough so that $\pi(\theta)$ is approximately constant for $\theta \in B$. Then

$$f(y \mid z, \theta)\pi(\theta) \approx f(y \mid z, \theta)\pi(\hat{\theta})$$

for all $\theta \in \Theta$. A quadratic log likelihood combines with a prior that is dominated by the data to give the following approximation:

$$\pi(\theta \mid z, y) \propto \exp[-\frac{1}{2}(\theta - \hat{\theta})'(-H(\hat{\theta}))(\theta - \hat{\theta})].$$

(The proportionality "constant" includes $\exp[L(\hat{\theta})]$ and $\pi(\hat{\theta})$.) This is the density of a multivariate normal distribution with mean $\hat{\theta}$ and covariance matrix equal to $-[H(\hat{\theta})]^{-1}$:

$$\theta^* \mid z, y \overset{a}{\sim} \mathcal{N}(\hat{\theta}, -[H(\hat{\theta})]^{-1}).$$

(A precise statement of this result is in the Bernstein-von Mises Theorem: *A Course in Large Sample Theory*, Thomas Ferguson, Chapman & Hall, 1996, Chapter 21; *Asymptotic Statistics*, A.W. van der Vaart, Cambridge University Press, 1998, Chapter 10.)

If we are interested in a linear combination $l'\theta^*$, then this result can be used to provide an approximate conditional (posterior) .95 interval:

$$\Pr\{l'\theta^* \in [l'\hat{\theta} - 1.96 \cdot \text{SE}, l'\hat{\theta} + 1.96 \cdot \text{SE}] \mid z, y\} \approx .95, \tag{2}$$

where

$$\text{SE} = (-l'[H(\hat{\theta})]^{-1}l)^{1/2}.$$

(a) Suppose that we have random sampling, so that (Y_i, Z_i) i.i.d. for $i = 1, \ldots, n$. Then we can apply the Note 13 results to obtain a limit distribution for $\sqrt{n}(\hat{\theta} - \theta^*)$, form an asymptotic pivot, and obtain an approximate .95 confidence interval for $l'\theta^*$. Is there a way to do this so that the interval we obtain is numerically equal to the approximate Bayesian .95 interval in (2)?

(b) Provide a brief discussion of how the interpretation of the frequentist .95 confidence interval differs from that of the Bayesian .95 posterior interval.

(c) Totrep needs to make a decision. (Trade-Off Talking Rational Economic Person, from *Notes on the Theory of Choice* by David Kreps, Westview Press, 1988.) There is a profit function, $g(a, \theta^*)$, which depends upon Totrep's action a and on the value of θ^* in (1). Totrep does not know θ^* but does have access to the data (Y_i, Z_i) for $i = 1, \ldots, n$ and would like to maximize expected profit. How might Totrep use the frequentist theory for maximum likelihood developed in Note 13? How might Totrep use the Bayesian approximation developed above?

2. Consider the panel probit model in Note 13:

$$\tilde{Y}_{it}^* = X_{it}'\alpha + \tilde{V}_i + \tilde{\epsilon}_{it},$$

with $\tilde{V}_i$, $\tilde{\epsilon}_{i1}$, ..., $\tilde{\epsilon}_{iT}$ mutually independent, $\tilde{V}_{it} \sim \mathcal{N}(0, \tilde{\sigma}_v^2)$, and $\tilde{\epsilon}_{it} \sim \mathcal{N}(0,1)$. (We have divided the original latent variable Y_{it}^* by σ_ϵ.) We observe

$$Y_{it} = \begin{cases} 1, & \text{if } \tilde{Y}_{it}^* \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

(a) Show that

$$\Pr(Y_{it} = 1 \,|\, X_{it}) = \Phi(X_{it}'\delta)$$

and provide a formula for δ.

(b) Show that there is a certain predictive effect that can be evaluated just using δ, without needing α and $\tilde{\sigma}_v$.

(c) Continue to assume random sampling on i. Try to generalize the latent variable model as much as you can, but keeping the implication that for some value of δ:

$$\Pr(Y_{it} = 1 \,|\, X_{it}) = \Phi(X_{it}'\delta) \qquad (t = 1, \ldots, T). \tag{$*$}$$

(The formula connecting δ to α and $\tilde{\sigma}_v$ will no longer hold.)

(d) Assume only that there is random sampling on i and that there is a value for δ such that $(*)$ holds. Consider applying a maximum-likelihood probit program to the cross-section observations for period t:

$$\hat{\delta}_t = \arg\max_a \sum_{i=1}^N \Big(Y_{it} \log[\Phi(X_{it}'a)] + (1 - Y_{it}) \log[1 - \Phi(X_{it}'a)] \Big).$$

Do this separately for $t = 1, \ldots, T$. Explain why each of these $\hat{\delta}_t$ is a consistent (as $N \to \infty$) estimator for δ.

(e) Stack the estimates in (d) into $\hat{\gamma}$ with $\hat{\gamma} \xrightarrow{p} \gamma$:

$$\hat{\gamma} = \begin{pmatrix} \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_T \end{pmatrix}, \quad \gamma = \begin{pmatrix} \delta \\ \vdots \\ \delta \end{pmatrix}.$$

Show that there is a moment function ψ that satisfies the key condition and

$$\sum_{i=1}^N \psi((Y_i, X_i), \hat{\gamma}) = 0.$$

Conclude that

$$\sqrt{N}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, \Lambda)$$

and provide a formula for Λ.

(f) Use the minimum-distance framework in Note 12 to combine the T separate estimates of δ and provide a confidence interval for $l'\delta$ (a linear combination of the coefficients). Provide enough detail so that a research assistant could program the procedure in Matlab. The research assistant is happy to take derivatives, but you need to be clear on what the function is and which derivatives are needed.

3. Suppose that Z_1 and Z_2 are binary random variables: Z_1 takes on only the values 0 and 1, Z_2 takes on only the values 0 and 1, and

$$0 < \text{Prob}\{Z_1 = l, Z_2 = m\} < 1 \qquad (l, m = 0, 1).$$

Consider the (population) linear predictor of Y given $1, Z_1, Z_2, Z_1 \cdot Z_2$:

$$E^*(Y \mid 1, Z_1, Z_2, Z_1 \cdot Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 \cdot Z_2.$$

(a) Does

$$E(Y \mid Z_1, Z_2) = E^*(Y \mid 1, Z_1, Z_2, Z_1 \cdot Z_2)?$$

Explain.

(b) Suppose that data (Y_i, Z_{i1}, Z_{i2}) are available from a random sample of $i = 1, \ldots, n$ individuals. The following four sample means have been tabulated:

$$\bar{Y}_{00}, \quad \bar{Y}_{01}, \quad \bar{Y}_{10}, \quad \bar{Y}_{11},$$

where

$$\bar{Y}_{lm} = \frac{\sum_{i=1}^n Y_i 1(Z_{i1} = l, Z_{i2} = m)}{\sum_{i=1}^n 1(Z_{i1} = l, Z_{i2} = m)} \qquad (l, m = 0, 1).$$

($1(B)$ is the indicator function that equals 1 if the event B occurs and equals 0 otherwise.) Use these means to provide an estimate of β_3. Is this a consistent estimator of β_3 as $n \to \infty$? Explain.

(c) The following four sample variances have been tabulated:

$$S_{00}, S_{01}, S_{10}, S_{11},$$

where

$$S_{lm} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_{lm})^2 1(Z_{i1} = l, Z_{i2} = m)}{\sum_{i=1}^n 1(Z_{i1} = l, Z_{i2} = m)},$$

and the sample sizes

$$n_{00}, n_{01}, n_{10}, n_{11},$$

where

$$n_{lm} = \sum_{i=1}^n 1(Z_{i1} = l, Z_{i2} = m) \qquad (l, m = 0, 1).$$

Let $\hat{\beta}_3$ denote your estimator for β_3 in (b). Use $\hat{\beta}_3$ along with these sample variances and sample sizes to provide a (approximate) .95 confidence interval for β_3. Does the probability that the interval covers β_3 converge to .95 as $n \to \infty$? Explain. Do not make additional assumptions such as homoskedasticity.

FINAL REVIEW PROBLEMS 4

1. In the normal linear model, the least squares estimator b is independent of the sum of squared residuals SSR conditional on $X = x$. What do we use this for?

2. If the $n \times 1$ random vector W has the (multivariate) normal distribution $\mathcal{N}(\mu, \Sigma)$, then what is the distribution of the scalar random variable $\alpha' W$ (where α is $n \times 1$ and not random)?

3. In the normal linear model, what is the distribution of the least-squares estimator? What is the distribution of the sum of squared residuals?

4. In the normal linear model, how do we construct a confidence interval for a linear combination of the coefficients?

5. Give an example where we would be interested in a confidence region for two or more linear combinations of the coefficients. Under the normal linear model, how do you determine whether a given point is in such a confidence region?

6. Suppose that you have a way to take independent draws from a standard normal distribution. How could you use this to obtain draws from a chi-square distribution? From a t distribution? From an F distribution?

7. Show that multiplication by an orthogonal matrix preserves the least-squares inner product.

8. How is the QR decomposition used to simplify derivations in the normal linear model?

9. Suppose that we have observations from a random sample, (Y_i, X_i) for $i = 1, \ldots, n$, where Y_i is scalar and X_i is $K \times 1$. Let $E^*(Y_i \mid X_i) = X_i' \beta$. We have been able to obtain a confidence interval for a linear combination $l' \beta$ such that the coverage probability converges to .95 as $n \to \infty$. Briefly discuss how we have been able to relax assumptions in the normal linear model in doing this. How does the confidence interval differ from the one in the normal linear model?

10. Suppose that we have observations from a random sample, (Y_i, X_i) for $i = 1, \ldots, n$, where Y_i is $H \times 1$ and X_i is $H \times K$.

(a) Provide a definition of the linear predictor $E^*(Y_i \mid X_i) = X_i \beta$. (Note that Y_i is a vector.) Provide a formula for β.

(b) Show that the least-squares estimator of β is consistent. What assumptions are needed?

(c) Provide explicit instructions for a research assistant to construct a .95 confidence

interval for a linear combination $l'\beta$. What assumptions are needed in order for the coverage probability to converge to .95?

11. Consider longitudinal data on a random sample of N firms followed over T periods.

(a) Provide minimal assumptions in order for the Mundlak estimator to have a limit distribution as $N \to \infty$.

(b) Provide minimal assumptions in order for the first-differences estimator to have a limit distribution as $N \to \infty$.

(c) Provide assumptions under which the Mundlak estimator and the first-differences estimator have the same probability limit as $N \to \infty$.

12. We have developed method of moments inference that exploits the orthogonality conditions in the following framework: $Q_i = R_i\gamma + V_i$, $E(W_iV_i) = 0$. Suppose that we have longitudinal data on earnings for a random sample of n individuals: $(Y_{i1}, \ldots, Y_{iT})$ $(i = 1, \ldots, n)$. Consider the following autoregression model:

$$E(Y_{it} \mid Y_{i1}, \ldots, Y_{i,t-1}, A_i) = \lambda + \gamma Y_{i,t-1} + A_i$$

$(t = 2, \ldots, T)$.

(a) Explain how the method of moments framework can be used to provide a (asymptotic, as $n \to \infty$) confidence interval for γ.

(b) Suppose that $T = 3$. Provide a consistent estimator for γ that is a ratio of sample covariances.

(c) Suppose that $T = 4$. Explain how a weight matrix is used to deal with the overidentification that results from there being more than one orthogonality condition for the estimation of the scalar parameter γ.

(d) Provide enough detail on the construction of an optimal weight matrix so that a research assistant, who knows Matlab, could use the data to do the calculations.

(e) Consider the weight matrix that would be optimal under homoskedasticity (so depends only on second moments, not fourth moments). Provide enough detail on the construction of this weight matrix so that a research assistant, who knows Matlab, could use the data to do the calculations. Can this weight matrix be used to obtain a confidence interval for γ that is valid in large samples even if there is heteroskedasticity? Explain.

13. Suppose that we have observations from a random sample, (Y_i, Z_i) for $i = 1, \ldots, n$, where Y_i is a binary random variable that takes on the values 0 and 1. The $K \times 1$ vector X_i is obtained from a function of Z_i that we specify: $X_i = f(Z_i)$.

(a) How is the probit approximation (based on X_i) to the regression function $E(Y_i \mid Z_i)$ defined?

(b) Sketch the argument that nonlinear least squares provides a consistent estimator for the coefficients in the probit approximation.

(c) Show how the GMM results can be used to obtain the limit distribution for the estimator in (b).

14. How is the information inequality used to provide intuition for the consistency of the maximum-likelihood estimator?

15. Explain how the GMM results can be used to obtain the limit distribution for the maximum-likelihood estimator.

16. Consider the following panel probit model:

$$Y_{it}^* = X_{it}'\alpha + V_i + \epsilon_{it},$$

with V_i, ϵ_{i1}, ..., ϵ_{iT} mutually independent, $V_{it} \sim \mathcal{N}(0, \sigma_v^2)$, and $\epsilon_{it} \sim \mathcal{N}(0, 1)$. We observe

$$Y_{it} = \begin{cases} 1, & \text{if } Y_{it}^* \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that you have maximum-likelihood estimates of α and σ_v. Discuss how they can be used to construct an estimate of the (predictive) effect of X_{it} on Y_{it}.

17. Suppose that you have observations from a random sample, (Y_{i1}, Y_{i2}) for $i = 1, \ldots, n$, where Y_{i1} and Y_{i2} are scalar. Consider the 3×1 vector $\hat{\sigma}$ formed from the sample variance of Y_1, the sample covariance between Y_1 and Y_2, and the sample variance of Y_2.

(a) Provide the limit distribution of $\hat{s}$.

(b) Provide a consistent estimator for the asymptotic covariance matrix of $\hat{s}$. Give enough detail so that a research assistant, who knows Matlab, could use the data to do the calculations.

18. You have the data from a random sample, Y_i for $i = 1, \ldots, n$, where the vector Y_i is 3×1: $Y_i' = (\, Y_{i1} \quad Y_{i2} \quad Y_{i3} \,)$. Let $\hat{\sigma}$ denote the 6×1 vector of distinct sample covariances, formed from the lower triangle of the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})'.$$

Consider the following variance-components model:

$$Y_{it} = \mu_t + V_i + U_{it} \qquad (t = 1, 2, 3),$$

where the latent random variables V_i, U_{i1}, U_{i2}, U_{i3} are all mutually independent with mean 0. Furthermore, the variance of U_{it} is constant:

$$\text{Var}(U_{it}) = \sigma_U^2 \qquad (t = 1, 2, 3).$$

Let σ_V^2 denote the variance of V_i. (μ_t is a parameter, giving the mean of Y_{it}.)

(a) Explain how the minimum distance framework can be applied to obtain a confidence interval for σ_V^2.

(b) Explain how to obtain a confidence interval for the variance ratio

$$\rho = \frac{\sigma_V^2}{\sigma_V^2 + \sigma_U^2}.$$

Provide enough detail so that a research assistant, who knows Matlab, could use the data to do the calculation.

(c) Consider the weight matrix for minimum distance that would be optimal if $Y_i \sim \mathcal{N}(\mu, \Sigma)$. Could this weight matrix be constructed just using the statistics $\bar{Y}$ and $\hat{\sigma}$? Explain. (It is not necessary to provide an explicit formula for this weight matrix.)

19. Suppose that you have data from a random sample, (Y_i, X_i) for $i = 1, \ldots, n$, where Y_i is scalar and X_i is $K \times 1$. Assume that

$$E(Y_i \,|\, X_i) = X_i'\beta, \quad \text{Var}(Y_i \,|\, X_i) = \exp(X_i'\gamma).$$

Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix}, \quad b = (X'X)^{-1}X'Y, \quad e = Y - Xb.$$

Consider the following nonlinear least-squares estimator for γ:

$$\hat{\gamma} = \arg\min_a \sum_{i=1}^n [e_i^2 - \exp(X_i'a)]^2.$$

(a) Explain how the GMM framework can be used to obtain a limit distribution for $\hat{\gamma}$.

(b) If β were known, we could use $U_i^2 = (Y_i - X_i'\beta)^2$ in the nonlinear least-squares estimator:

$$\hat{\gamma} = \arg\min_a \sum_{i=1}^n [U_i^2 - \exp(X_i'a)]^2.$$

Does this estimator have the same limit distribution as the estimator in (a) that uses e_i^2? Explain.

20. How can an instrumental variable model be used to deal with omitted variable bias? What are the requirements on the instrumental variable? How can the orthogonality condition ($\perp$) framework be applied to obtain inferences?

21. In what sense is two-stage least squares an optimal estimator?

22. Consider a job training program that is intended to raise the future earnings of the participants. We want to evaluate the effectiveness of the program. The target population consists of high school dropouts. A random sample of size $n = 5000$ is taken from this population and everyone in the sample is invited to participate in the program, which takes six months. In addition, 2500 of the sample members are chosen at random and offered a subsidy of \$500 per month if they participate in the program. For each of the n sample members, we observe B, S, T, and Y: B = earnings for the individual in 1970, before the program took place; $S = 1$ if the individual is offered the subsidy, $S = 0$ otherwise; $T = 1$ if the individual enrolls in the program, $T = 0$ otherwise; Y = earnings for the individual in 1975, two years after the program took place. Suppose that everyone who enrolls in the program completes it.

(a) Define a treatment effect of the program on earnings. You may assume that the treatment effect is the same for all the individuals.

(b) Consider two least-squares regressions: (i) Y on a constant and T; (ii) Y on a constant, T, and B. Which of these regressions would be better for estimating the treatment effect in (a)? Explain your reasoning. Under what conditions would the preferred regression provide a consistent estimate of the treatment effect?

(c) Explain how you could use the information on the subsidy to construct an instrumental variables estimator of the treatment effect. Sketch the argument that motivates this estimator. What are the advantages and disadvantages of this estimator relative to your preferred regression estimator in (b)?

23. Explain how random assignment deals with omitted variable bias. Explain how panel data can be used to deal with omitted variable bias. What are the advantages and disadvantages of these two approaches to dealing with omitted variable bias?

24. Consider estimating the demand function in the market for Frozen Orange Juice Concentrate (FOJC) in the U.S. There are n annual observations on market price (p) and quantity (q). In addition, you have data on annual Florida rainfall in inches (r) and on U.S. per capita income (m).

(a) A least-squares regression of log q on log p gives a slope coefficient b. Is b likely to be biased upward or downward as an estimate of the demand elasticity? Explain.

(b) Build a model that incorporates all the available data (i.e., p, q, r, and m). Explain the role played by each of the variables and the assumptions you are using.

(c) Show how the model in (b) motivates a two-stage least squares estimator for the demand elasticity.