Economics 696, Topics in Econometrics

Lecture Note 9: Panel Data

1 Cobb-Douglas Production Function

To fix ideas we will consider a simple example: estimating a Cobb-Douglas production function. We will start by going through some standard panel data techniques at a heuristic level before considering the probabilistic structure and identification more formally.

1.1 Cross-sectional Analysis

Consider a firm which produces output using a technology described by a Cobb-Douglas production function:

$$Y(K, L) = AK^{\gamma_1}L^{\gamma_2}. \tag{1}$$

Here K is capital input, L is labor input, and Y is output. If $\gamma_1 + \gamma_2 < 1$ (so there are decreasing returns to scale), we might suppose that the firm chooses capital and labor to maximize profits, taking output price p_y and input prices p_k, p_l as given. In this case, factor demands for K and L will be increasing in A, since firms will hire inputs until their marginal revenue product equals their price.

We are interested in connecting this economic model to empirical data on firms. We will be working with data based on many firms, so we need to specify the production functions for *each* firm. Suppose we can write, for $i = 1, \ldots, n$,

$$Y_i(K, L) = A_i K^{\gamma_1} L^{\gamma_2}.$$

This expresses the notion that each firm has a Cobb-Douglas production function with common coefficients γ_1, γ_2. However, there are differences in how efficient the firms are, which arises from variation in A_i across firms.

Here, K, L are not "data" but are simply arguments in the function $Y_i(K, L)$. We will use K_i and L_i to denote the amounts of capital and labor actually chosen by the firm, and Y_i to denote the actual output of the firm. Then the Cobb-Douglas model implies that

$$Y_i = A_i K_i^{\gamma_1} L_i^{\gamma_2},$$

or taking logs:

$$\log Y_i = \log A_i + \gamma_1 \log K_i + \gamma_2 \log L_i.$$

To simplify notation, let us define $y_i \equiv \log Y_i$, and similarly for A_i, K_i, and L_i. Then we can write

$$y_i = b + \gamma_1 k_i + \gamma_2 l_i + u_i,$$

where $b \equiv E(a_i)$, and $u_i = a_i - b$. This looks like a classical regression model for y_i given k_i and l_i. However, in the classical regression model, the disturbance is assumed to satisfy

$$E(u_i | k_i, l_i) = 0.$$

So we would be assuming that u_i is mean-independent of k_i and l_i. But recall that under price-taking and profit-maximization, we would expect that k_i and l_i are related quite strongly to efficiency a_i and hence to u_i. So OLS does not seem to be an appropriate estimator here. (As noted above, we will defer considering formal properties of estimators while we establish some concepts and notation.)

1.2 The Role of Panel Data

One possible solution emerges if the firms are observed in multiple time periods. Suppose for each firm, we observe output and measured inputs in each of T years. We will denote these observations by (y_{it}, k_{it}, l_{it}), for $i = 1, \ldots, n$, $t = 1, \ldots, T$. This is an example of *panel data*. In general, the term panel data refers to any data with a natural grouping structure. Another example of panel data is data on earnings and other variables for each sibling in a family, for a large number of families. Suppose that our previous model continues to hold, so that

$$y_{it} = a_{it} + \gamma_1 k_{it} + \gamma_2 l_{it}.$$

Here a_{it} is interpreted as a measure of firm i's efficiency at time t. If we write $a_{it} \equiv \alpha_i + u_{it}$, then we can write our model as

$$y_{it} = \alpha_i + \gamma_1 k_{it} + \gamma_2 l_{it} + u_{it}.$$

We could interpret α_i as capturing firm-specific inputs, such as management quality, which do not change over time. We might then assume that

$$E(u_{it}|l_i, k_i, \alpha_1, \ldots, \alpha_n) = 0.$$

Here we use $l_i = (l_{i1}, \ldots, l_{iT})$ and $k_i = (k_{i1}, \ldots, k_{iT})$. The model looks like a classical regression model, except: (1) there is a different intercept term for each firm; and (2) the conditioning variables are a little different. The connection is even stronger if we define dummy variables

$$d_{it,j} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{otherwise.} \end{cases}$$

Write $x_{it} \equiv (k_{it}, l_{it})'$, $d_{it} \equiv (d_{it,1}, \ldots, d_{it,n})'$, $\gamma \equiv (\gamma_1, \gamma_2)'$ and $\alpha = (\alpha_1, \ldots, \alpha_n)$. Then

$$E(y_{it}|x_i, \alpha) = x_{it}'\gamma + d_{it}'\alpha.$$

It is important to note the rather strong assumptions that we have made use of so far. The strongest is that only the *constant* part of unmeasured efficiency $a_{it} = \alpha_i + u_{it}$ is relevant for the firm's choice of k_{it} and l_{it}. This might be reasonable if u_{it} represents factors which cannot be predicted at the time the firm chooses its factor inputs. For example, if we are looking at farms' production of some crop, this could be affected by changes in weather which are not anticipated at the time the inputs are chosen. But if this assumption is not reasonable, then it may be difficult to estimate the parameters of the production function reliably.

2 Fixed Effects

Our model is

$$E(y_{it}|X,\alpha) = x_{it}'\gamma + d_{it}'\alpha, \tag{2}$$

where x_{it} is a $k \times 1$ vector of regressors (which does not include a constant), and d_{it} is a $n \times 1$ vector of dummy variables as defined above, and X is interpreted to contain all the regressors and the dummy variables. Let $\beta = (\gamma', \alpha')'$, and

$$X = \begin{pmatrix} x_{11}' & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1T}' & 1 & 0 & \cdots & 0 \\ x_{21}' & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ x_{2T}' & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n1}' & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ x_{nT}' & 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{n1} \\ \vdots \\ y_{nT} \end{pmatrix}$$

The least squares estimator is

$$\hat{\beta} = \begin{pmatrix} \hat{\gamma} \\ \hat{\alpha} \end{pmatrix} = (X'X)^{-1}X'y,$$

with corresponding variance estimator

$$s^2 = \frac{1}{nT-k-n}(y - X\hat{\beta})'(y - X\hat{\beta}) = \frac{1}{nT-k-n}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - x_{it}'\hat{\gamma} - \hat{\alpha}_i)^2.$$

NOTE: we have not yet shown that the least-squares estimator has any particular statistical properties (unbiasedness, consistency, etc.) We cannot actually do so yet, because we have not formally defined a sampling mechanism. In particular, we need to be careful about how we treat the α_i terms.

The least-squares estimate $\hat{\beta}$ is often called the "fixed effects" estimator, or FE for short. Another name is the least-square dummy variables (LSDV) estimator. It is typically the case that n, the cross-sectional dimension, is large relative to T. In this case, X can contain a large number of columns because there are many dummy variables. This means that $X'X$ is a large matrix, possibly difficult to invert on a computer with limited memory. The following results can be used to simplify the calculations:

Result 1

$$\hat{\gamma} = (X_w'X_w)^{-1}X_w'y_w,$$

and

$$Var(\hat{\gamma}) = \sigma^2 (X_w' X_w)^{-1},$$

where

$$X_w = \begin{pmatrix} (x_{11} - \bar{x}_1)' \\ \vdots \\ (x_{1T} - \bar{x}_1)' \\ \vdots \\ (x_{n1} - \bar{x}_n)' \\ \vdots \\ (x_{nT} - \bar{x}_n)' \end{pmatrix}, \qquad y_w = \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1T} - \bar{y}_1 \\ \vdots \\ y_{n1} - \bar{y}_n \\ \vdots \\ y_{nT} - \bar{y}_n \end{pmatrix},$$

$$\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}, \qquad \bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}.$$

Stated in this form, the estimator is often called the *within* estimator, because it is based on deviations from within-firm averages. The interpretation of this result is that in order to get the least-squares estimates for γ, one can perform the shorter regression given above. To obtain an estimate of the variance, the following is useful:

Result 2:

$$y_{it} - x_{it}'\hat{\gamma} - \hat{\alpha}_i = (y_{it} - \bar{y}_i) - (x_{it} - \bar{x}_i)'\hat{\gamma}.$$

Thus

$$s^2 = \frac{1}{nT - k - n}(y_w - X_w\hat{\gamma})'(y_w - X_w\hat{\gamma}).$$

Notice that this is different from the variance estimate that a "canned" least-squares routine applied to the previous short regression would produce.

3 Random Effects

There is an alternative approach to working with panel data models which connects nicely to GLS estimation. For simplicity assume all variables are measured in deviations from (grand) means. (Otherwise we could let the vector x_{it} include a constant and all of what follows would go through.) Assume that the α_i are i.i.d. with

$$E(\alpha_i|X) = 0, \qquad V(\alpha_i|X) = \sigma_\alpha^2.$$

Defining $\epsilon_{it} = \alpha_i + u_{it}$, we can write the model as

$$y_{it} = x_{it}'\gamma + \epsilon_{it}, \qquad i = 1, \ldots, n; t = 1, \ldots, T.$$

Stacking the observations:

$$y = X_1\gamma + \epsilon,$$

where y is as defined before and

$$X_1 = \begin{pmatrix} x'_{11} \\ \vdots \\ x'_{1T} \\ \vdots \\ x'_{n1} \\ \vdots \\ x'_{nT} \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \alpha_1 + u_{11} \\ \vdots \\ \alpha_1 + u_{1T} \\ \vdots \\ \alpha_n + u_{n1} \\ \vdots \\ \alpha_n + u_{nT} \end{pmatrix} = \begin{pmatrix} \alpha_1 \iota \\ \vdots \\ \alpha_n \iota \end{pmatrix} + u.$$

Here ι is a $T \times 1$ vector of ones. We can write

$$E(\epsilon|X) = 0,$$

and

$$V(\epsilon|X) = \begin{pmatrix} \sigma_\alpha^2 \iota\iota' & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_\alpha^2 \iota\iota' \end{pmatrix} + \sigma^2 I_{nT} \equiv \Omega.$$

We can think of this as an "error components" model in that there is a composite error term arising from the α_i and the u_{it}; this gives the variance matrix a particular correlation structure. In any case, now we have a generalized regression model with variance matrix Ω. It would be natural to apply GLS. This leads to the estimator

$$\hat{\gamma}_{GLS} = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y.$$

The GLS estimator has variance

$$V(\hat{\gamma}_{GLS}|X) = (X_1'\Omega^{-1}X_1)^{-1}.$$

3.1 Within and Between

The GLS estimator in this model is called the random effects (RE) estimator. It can be shown to have the following form, which leads to some additional insight.

Result 3:

$$\hat{\gamma}_{GLS} = (X_w'X_w + rX_b'X_b)^{-1}(X_w'X_w\hat{\gamma}_w + rX_b'X_b\hat{\gamma}_b),$$

where X_w and y_w are the deviations from means as defined before, and

$$X_b = \begin{pmatrix} \bar{x}_1' \\ \vdots \\ \bar{x}_1{}' \\ \vdots \\ \bar{x}_n' \\ \vdots \\ \bar{x}_n' \end{pmatrix}, \qquad y_b = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_1 \\ \vdots \\ \bar{y}_n \\ \vdots \\ \bar{y}_n \end{pmatrix},$$

$$\hat{\gamma}_w = (X_w' X_w)^{-1} X_w' y_w, \qquad \hat{\gamma}_b = (X_b' X_b)^{-1} X_b' y_b,$$

and

$$r = (1 + T \frac{\sigma_\alpha^2}{\sigma^2})^{-1}.$$

Notice that $\hat{\gamma}_w$ is just the fixed effects (within) estimator. We will call $\hat{\gamma}_b$ the "between" estimator because it only uses variation between units. The RE estimator is a weighted combination of the within and between estimates, where the weight depends on the relative variances of the individual "effect" α_i and the idiosyncratic disturbance u_{it}. As $\sigma_\alpha^2 \to \infty$, $r \to 0$ and we obtain the FE estimator.

3.2 Feasible GLS in the Random Effects Model

In order to implement feasible GLS in the RE model, we only need estimates of σ^2 and σ_α^2. With these we can form an estimate of Ω; alternatively, we can form an estimate of r and use the between and within estimates as in Result 3.

An easy way to estimate σ^2 is to just use the variance estimate $\hat{\sigma}^2$ arising from the FE estimator. Based on Result 2, this can be obtained using the within estimator.

To estimate σ_α^2, we can use the between estimates. Notice that the equation being estimated by the between estimator is

$$\bar{y}_i = \bar{x}_i{}' \gamma + v_i,$$

where

$$v_i \equiv \alpha_i + \bar{u}_i.$$

The variance of the disturbance term is

$$\sigma_v^2 = \sigma_\alpha^2 + \frac{\sigma^2}{T}.$$

Thus, if we use the residuals from the between estimates to form a variance estimate $\hat{\sigma_v}^2$, we can obtain an estimate of σ_α^2 by

$$\hat{\sigma}_\alpha^2 = \hat{\sigma}_v^2 - \frac{\hat{\sigma}^2}{T}.$$

3.3 Comparing the FE and RE estimators

One might expect that the random effects estimator is superior to the fixed effects estimator. After all, it is the GLS estimator; moreover, the previous discussion shows that the fixed effects estimator is a limiting case of RE, corresponding to situations where the variation in the individual effects is large. Since the feasible version can actually estimate the variance of the individual effects, this would seem preferable to assuming it is arbitrarily large. However, there is a very strong assumption built in to the random effects estimator: this is the assumption that the disturbances, including α_i, are orthogonal to the explanatory variables. Going back to our production function example, this was exactly the case we wanted to avoid. So the RE estimator may not be appropriate for that case; in other applications where the omitted variables interpretation of α_i is less relevant, this may be less of an issue.

4 FE Model: Probability Structure, Identification, and Asymptotics

The random effects model of the previous section specified a joint distribution for the error components $\alpha_i, u_{i1}, \dots, u_{iT}$ conditional on X. This made the model sufficiently "complete" so that the statistical properties of estimators like GLS are well defined. However, it had the drawback that the random effects assumption was inconsistent with our original story of firm managers choosing inputs to maximize expected profits.

The FE estimator appears to make much weaker assumptions, because it does not specify how α_i is related to the factor inputs. Intuitively, by treating the α_i as if they were parameters to be estimated, it seems to sidestep the issue of specifying a dependence structure between the α_i and the factor inputs. But this raises a number of questions: how do we interpret the α_i, as parameters or latent variables (or both)? And more generally, what is the underlying probabilistic structure generating this doubly-indexed data? This was never fully resolved in Sections 1 and 2. Without saying a bit more about these issues, it is not clear what we even mean by identification.

Here is one way to set up the probabilistic structure of the model. Suppose that n (the number of firms) is large relative to T (the number of time periods). Then it would be natural to think of randomly sampling firms i from the population of firms, where each "draw" is a short time series ($t = 1, \dots, T$). Formally, let

$$\alpha_i \overset{\text{iid}}{\sim} F_\alpha,$$

and let the u_{it} be drawn independently of each other and of α_i with mean 0:

$$u_{it} | \alpha_i, u_{i1}, \dots, u_{i,t-1}, u_{i,t+1}, \dots, u_{iT} \overset{\text{iid}}{\sim} F_u, \quad \text{where } F_u \text{ has mean } 0.$$

(We could relax this independence somewhat, but it is important that the u_{it} are orthogonal to α_i.) Thus we have specified a joint distribution for $(\alpha_i, u_{i1}, \dots, u_{iT})$.

Now consider the factor inputs x_{it}. Per our discussion above, the firms should make factor input choices based on α_i but not the u_{it}. Assume that $x_i = (x_{i1}, \dots, x_{iT})$ are independent across firms with conditional joint CDF

$$F_{x|\alpha,u}(x_{i1}, \dots, x_{iT} | \alpha_i, u_{i1}, \dots, u_{iT}) = F_{x|\alpha}(x_{i1}, \dots, x_{iT} | \alpha_i).$$

Finally, the outputs are determined by

$$y_{it} = \alpha_i + x_{it}'\gamma + u_{it}.$$

This generates (for a given $F_{\alpha,u}$, $F_{x|\alpha}$, and γ) the observable joint distribution $F_{y,x}$ of $y_{i1}, \ldots, y_{iT}$, $x_{i1}, \ldots, x_{iT}$. We do *not* observe the latent α_i (nor the u_{it}).

This setup captures our original story about firms' factor input choices, because it lets the factor choice depend on α_i. However, it seems somewhat at odds with the FE estimator, which treats the α_i as if they were parameters. But since the FE estimator at this point is merely a procedure for turning the data into an estimate of γ, we can examine the properties of FE under our probabilistic assumptions.

From the specification of y_{it}, we can write

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\gamma + (u_{it} - \bar{u}_i).$$

Notice that the α_i terms drop out as a consequence of taking within-firm deviations. Also, we have that $E[u_{it} - \bar{u}_i | x_{i1}, \ldots, x_{iT}] = 0$, since the u_{it} have mean 0 and are independent of the x_{it}. So for $t = 1, \ldots, T$,

$$E[y_{it} - \bar{y}_i | x_{i1}, \ldots, x_{iT}] = (x_{it} - \bar{x}_i)'\gamma.$$

Hence γ is identified from the observable distribution $F_{y,x}$.

In fact, the FE estimator works off of this identification result. Consider the within estimator representation of $\hat{\gamma}$. We can write

$$\begin{aligned} \hat{\gamma} &= (X_w' X_w)^{-1} X_w' y_w \\ &= \left(\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i). \end{aligned}$$

Therefore,

$$\hat{\gamma} = \gamma + \left(\sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(u_{it} - \bar{u}_i).$$

The rightmost term converges to zero in probability as $n \to \infty$ since $E[(x_{it} - \bar{x}_i)(u_{it} - \bar{u}_i)] = 0$. So $\hat{\gamma}$ will be consistent for γ.

The FE estimator succeeds because it implicitly removes the role of the latent α_i through mean-differencing. An alternative would be to model the relationship between α_i and the factor inputs.

If we assume that the factor inputs are independent of α_i, then we get the random effects model of section 3. We could extend the random effects model to allow the α_i to be related to the x_i. For example, we could assume that

$$\alpha_i = \bar{x}_i' \pi + v_i,$$

where $E(v_i | x_{i1}, \ldots, x_{iT}) = 0$. Or we could assume a somewhat more general model, such as

$$\alpha_i = (x_{i1}', \ldots, x_{iT}') \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_T \end{pmatrix} + v_i.$$

This would build in correlation between α_i and the regressors, but in a limited way. For more on this approach, see Chamberlain (1984).

5 Problems with the Dummy Variable Approach

At this point, we might conclude that the dummy variable (FE) approach is very attractive. It is easy to implement, and we do not have to make *any* assumptions about the α_i. Although the α_i themselves cannot be consistently estimated, we found we could estimate the γ consistently. Unfortunately, naive application of the dummy variable approach can lead to problems in many panel data models.

For the within estimator to be consistent, we needed

$$E\left[(x_{it} - \bar{x}_i)(u_{it} - \bar{u}_i)\right] = 0.$$

This holds if the x_{it} are "strictly exogenous" with respect to the entire sequence of disturbances, in other words if

$$E(x_{is}u_{it}) = 0, \qquad \forall t, s.$$

This assumption might be too strong. For example, suppose that factor inputs x_{it} do not respond immediately to innovations to efficiency u_{it}, but do respond to past innovations. Then we might only have the weaker condition that

$$E(x_{is}u_{it}) = 0, \qquad s \leq t,$$

which says that the x_{it} are "predetermined" but not strictly exogenous. Then the within estimator will not be consistent, since

$$E\left[(x_{it} - \bar{x}_i)(u_{it} - \bar{u}_i)\right] \neq 0.$$

Other problems arise when the equation being estimated is not linear. Then simply using individual/firm dummies may lead to problems. In many of these more difficult cases, it is possible to construct consistent estimators, often based on IV or GMM methods. One solution is based on a first-differencing of the data:

$$(y_{it} - y_{i,t-1}) = \gamma'(x_{it} - x_{i,t-1}) + (u_{it} - u_{i,t-1}).$$

Notice that, similar to the within transformation, first-differencing the data removes the individual/firm-specific effect α_i. Simply running a regression of $(y_{it} - y_{i,t-1})$ on $(x_{it} - x_{i,t-1})$ will not lead to a consistent estimator, however, since

$$E(x_{it}u_{i,t-1}) \neq 0.$$

But under the predeterminedness assumption, $x_{i,t-1}$ *is* orthogonal to both u_{it} and $u_{i,t-1}$. Thus

$$E(x_{i,t-1}(u_{it} - u_{i,t-1})) = 0,$$

so we can use $x_{i,t-1}$ as an instrument for $(x_{it} - x_{i,t-1})$. More generally, we have

$$E(x_{is}(u_{it} - u_{i,t-1})) = 0, \qquad s \leq t-1.$$

Therefore we can build an IV estimator based on the orthogonality conditions

$$E\left[x_{is}((y_{it} - y_{i,t-1}) - \gamma'(x_{it} - x_{i,t-1}))\right] = 0, \qquad s \leq t-1.$$

As another example of the inconsistency of the within estimator, consider the dynamic panel data model

$$y_{it} = \alpha_i + \rho y_{i,t-1} + u_{it}.$$

This relates the current value of y_{it} to its previous ("lagged") value as well as α_i. The parameter ρ captures the degree of autocorrelation in y_{it}. This kind of model is used extensively in studying individual earnings, as well as economic growth of countries and other inherently dynamic relationships.

Let us suppose that y_{i0} (the "initial condition") is observed for each i, and that the u_{it} are i.i.d., independent of $y_{i0}, \ldots, y_{i,t-1}$, with mean 0 and variance σ^2.

By setting $x_{it} \equiv y_{i,t-1}$, we see that the LSDV estimator for ρ would regress

$$y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

on

$$y_{i,t-1} - \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}.$$

The problem is that now, $(u_{it} - \bar{u}_i)$ will *not* be orthogonal to $(y_{i,t-1} - \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1})$, since by construction $y_{i,t-1}$ depends on $u_{i,t-1}$. As a consequence, $\hat{\rho}$ will not consistently estimate ρ. When T is very small (e.g. 2), the inconsistency can be quite severe.