University of Illinois Department of Economics
Fall 2014 Roger Koenker

Economics 508
Lecture 17 Panel Data

Consider a model of the form

$$y_{it} = x_{it}\beta + z_i\gamma + \alpha_i + u_{it} \quad i = 1, \ldots, n \;\; t = 1, \ldots T \tag{1}$$

where, for example,

$y_{it} =$ log wage of person i at time t.
$x_{it} =$ time varying characteristics at time t like age, experience, health, ...
$z_i =$ time invariant characteristics at time t like education, race, sex, ...
$\alpha_i =$ unobserved individual effect like spunk, ability
$u_{it} =$ everything else.

we will stack the model so that all T observations on person 1 comes first, and then person 2, and so on.

Now consider the matrix,

$$P = I_n \otimes T^{-1} 1_T 1_T^\top \equiv I_n \otimes J_T$$

where the latter matrix is T^{-1} times a matrix of ones. It is easy to see that P represents an orthogonal projection, it is symmetric and idempotent. What does it do? Consider

$$Py = \begin{bmatrix} J_T & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots & \\ & & & & J_T \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \bar{y}_1 1_T \\ \vdots \\ \vdots \\ \vdots \\ \bar{y}_n 1_T \end{bmatrix}$$

And therefore,

$$Qy \equiv (I - P)y = y - \bar{y}$$

is a deviation-from-individual-means vector. Note that if we wanted to view P as representing a least squares projection, we can think of it as arising from a model in which there are dummy variables for just the individual effects,

$$y_{it} = \alpha_i + u_{it}$$

We might write this as,

$$y = D\alpha + u.$$

It is a useful exercise to show that $P = D(D^\top D)^{-1} D^\top$ where $\hat{y} = Py$ would be the least squares fit and $\hat{u} = Qy$ would be the residual vector. Clearly applying Q to D yields,

$$QD = 0$$

since there is no temporal variability in D by hypothesis. A common estimator of (1) for at least the β component is

$$\hat{\beta}_W = (X^\top QX)^{-1}X^\top Qy$$

which is frequently called the "within group" estimator. As long as we assume

$$Ex_{it}u_{it} = 0$$

$\hat{\beta}_W$ is consistent for β. But as the name suggests, $\hat{\beta}_W$ uses only some of the information available. Note that this is all just the usual Gauss-Frisch-Waugh machinery in which we are removing the effect of D before getting down to the business of estimating the effects of X.

A small point that was raised in class is to note that if we multiply through the original model equation by Q, this annihilates the α's and the Z term, but it also transforms the error u to Qu. What is the consequence of this? If the original u is iid, say $\mathcal{N}(0, \sigma^2 I_{nT})$ then Qu is $\mathcal{N}(0, \sigma^2 Q)$ thus if we do GLS we seem to need to invert Q. Of course this isn't feasible, but if we instead use *any* g-inverse the estimator uses weights $QQ^- Q = Q$, by the requirement discussed in L13. So we have just the usual second step of the Gauss-Frisch-Waugh procedure.

We also have the "between groups" information which is obtained by multiplying (1) by P

$$\bar{y}_i = \bar{x}_i\beta + z_i\gamma + \alpha_i + \bar{u}_i$$

Note here that we can delete the $n(T-1)$ redundant observations. Let's denote OLS estimators of (β, γ) as $(\hat{\beta}_B, \hat{\gamma}_B)$ for "between." In this case we have only n observations so we can't possibly estimate the α_i's so we have to consider the composite error term, $\alpha_i + u_{it}$. Can we combine β_B, β_W somehow?

A Simple Measurement Error Problem (Revisited)

1. Suppose that $y_i \sim \mathcal{N}(\mu, \sigma_i^2)$ for $i = 1, 2$. The GLS estimator of μ is:

   $$\hat{\mu} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y$$

   where

   $$\Omega = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \qquad X = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

   so

   $$\hat{\mu} = (1/\sigma_1^2 + 1/\sigma_2^2)^{-1}[y_1/\sigma_1^2 + y_2/\sigma_2^2]$$

2. Matrix Case – interpret as two independent estimates

   $$y_i \sim \mathcal{N}_p(\mu, \Omega_i) \qquad i = 1, 2$$

   $$\hat{\mu} = (\Omega_1^{-1} + \Omega_2^{-1})^{-1}[\Omega_1^{-1}y_1 + \Omega_2^{-1}y_2]$$

   here

   $$\Omega = \begin{bmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{bmatrix} \quad X = \begin{bmatrix} I_p \\ I_p \end{bmatrix}$$

   Note that if we put in values for σ_2 or Ω_2, and let them tend to infinity, then we get just first component.

To apply this to get the expression at the top of page 1381 of *HT*, note that if

$$\hat{\beta}_i \sim \mathcal{N}(\beta, V_i) \qquad i = W, B$$

our simple weighted least squares approach gives,

$$\hat{\beta} = (V_B^{-1} + V_W^{-1})^{-1}[V_B^{-1}\hat{\beta}_B + V_W^{-1}\hat{\beta}_W]$$

HT rewrite this incorrectly so *be careful*! Note that

$$V_B^{-1} + V_W^{-1} = V_B^{-1}(V_B + V_W)V_W^{-1}$$

so

$$(V_B^{-1} + V_W^{-1})^{-1} = V_W(V_B + V_W)^{-1}V_B$$

so we may write,

$$\hat{\beta} = \Delta\hat{\beta}_B + (I - \Delta)\beta_W$$

where

$$\Delta = V_W(V_B + V_W)^{-1}$$

Note *HT* write $\Delta = (V_B + V_W)^{-1}V_W$! Exercise: Verify that $\hat{\beta}_B$ and $\hat{\beta}_W$ have covariance zero in order to justify the application of the above result.

Generalized Least Squares Now let's consider GLS estimation of the the somewhat simplified model, that ommits the time invariant covariates,

$$y = X\beta + D\alpha + u$$

By treating the α_i's as random and u_i and α_i as independent. we can compute the covariance matrix of $\epsilon = u + D\alpha$ as,

$$\begin{aligned}
\Omega = E\varepsilon\varepsilon^\top &= E(D\alpha + u)(D\alpha + u)^\top \\
&= \sigma_\alpha^2 DD^\top + \sigma_u^2 I_{nT} \\
&= \sigma_u^2 I_{Tn} + \sigma_\alpha^2(I_n \otimes TJ_T) \\
&= \sigma_u^2 I_{nT} + T\sigma_\alpha^2 P
\end{aligned}$$

Thus we may write the GLS estimator with $\tilde{X} = [X \vdots Z]$,

$$\hat{\delta} = \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = (\tilde{X}^\top\Omega^{-1}\tilde{X})^{-1}\tilde{X}^\top\Omega^{-1}y$$

This is sometimes called the Balestra-Nerlove estimator. Using the following Lemma it can be reformulated as yielding a simple least squares estimator after a preliminary transformation of the data.

Lemma (Nerlove) Let $\sigma_\epsilon^2 = \sigma_u^2 + T\sigma_\alpha^2$, then $\Omega^{-1/2} = \sigma_\epsilon^{-1}P + \sigma_u^{-1}Q$.

Proof: We will show this by computing directly that $\Omega^{-1/2}\Omega\Omega^{-1/2} = I_{nT}$, Noting that $PQ = 0$, we have,

$$\begin{aligned}(\sigma_\varepsilon^{-1}P + \sigma_u^{-1}Q)[\sigma_u^2 I + T\sigma_\alpha^2 P](\sigma_\varepsilon^{-1}P + \sigma_u^{-1}Q) &= \sigma_\varepsilon^{-2}(\sigma_u^2 + T\sigma_\alpha^2)P + \sigma_u^{-2}\sigma_u^2 Q \\ &= P + Q \\ &= I_{nT}\end{aligned}$$

Remark: Ω has only 2 distinct eigenvalues $\sigma_u^2 + T\sigma_\alpha^2$ and σ_u^2 and corresponding eigenvectors P and Q.

Having computed $\Omega^{-1/2}$ we can transform (1) by $\Omega^{-1/2}$ to obtain a spherical error, HT use $\sigma_u\Omega^{-1/2}$ to get.

$$\sigma_u\Omega^{-1/2}y = (\theta P + Q)y = y - (1-\theta)\bar{y}$$

where $\theta = \sigma_u/(\sigma_u^2 + T\sigma_\alpha^2)^{1/2}$, and similarly for the other variables. Here we are doing a form of "partial deviations from means" analogous to partial differencing in autocorrelation correction. Under our assumption, such estimates are efficient. Note that in unbalanced cases the transformations are slightly more complicated.

A third approach to deriving the classical panel data estimation methods for random effects is to regard the individual effects, α_i, as parameters that need to be estimated, but ones on which you have some prior information and procede in a Bayesian fashion. To this end, we can consider, minimizing a penalized log likelihood expression,

$$min_{\alpha,\beta,\gamma} \sum_i \sum_t (y_{it} - x_{it}\beta - z_i\gamma - \alpha_i)^2 + \lambda \sum_i \alpha_i^2.$$

where the λ represents the strength of the prior. The effect of the (second) penalty term is to shrink the fixed effect estimates – that would be obtained if the model were estimated with $\lambda = 0$ – toward zero. Ideally, one would choose $\lambda = \sigma_u^2/\sigma_\alpha^2$. It is not particularly obvious from the linear algebra, but it turns out that this approach is equivalent to the two prior approaches that have been described.

Specification Tests

Intuitively, if our assumption is violated, then β_W is still consistent for β, but inefficient relative to the optimal $\hat{\beta} = \Delta_{11}\hat{\beta}_B + (I - \Delta_{11})\hat{\beta}_W$. This seems to be ideally suited for the H-test. We have An efficient estimator under H_0 which is inconsistent under $H_A: \ \hat{\beta}$ and a consistent estimator under $H_A: \ \hat{\beta}_W$

There are three obvious options for testing: $\omega_1 = \hat{\beta} - \hat{\beta}_W$, $\omega_2 = \hat{\beta} - \hat{\beta}_B$, and $\omega_3 = \hat{\beta}_W - \hat{\beta}_B$

HT show that the three tests are asymptotically equivalent. As in other H-tests we can use the fact that under H_0, e.g., $V(\hat{\beta} - \hat{\beta}_W) = V(\hat{\beta}_{sW}) - V(\hat{\beta})$.

Estimation of γ. Recall that we still have problems with estimation of γ in the fixed effects model and we might want to use fixed effects if we believed that there were endogoneity problems. We can think of distinguishing

$$\begin{aligned} X &= [X_1 \vdots X_2] \\ Z &= [Z_1 \vdots Z_2] \end{aligned}$$

where as in *HT* X_2 and Z_2 are to be treated as *endogenous* and $[X_1 \vdots Z_1]$ as exogenous. Then we write

$$\tilde{y} = \tilde{X}\beta + \tilde{Z}\gamma + \tilde{\varepsilon}$$

where $\Omega^{-1/2}y = \tilde{y}$ and so forth, and we have the two reduced form equations

$$[X_2 \vdots Z_2] = [X_1 \vdots Z_1]\Pi$$

For slightly esoteric reasons 2SLS and 3SLS are equivalent here – basically because of the fact that the "other equations" are *exactly* identified.

A General Approach to Computation

The simplest, but perhaps not most memory efficient means of estimation is to take

$$\tilde{y} = \tilde{X}\beta + \tilde{Z}\gamma + \tilde{\varepsilon}$$

and define the instruments

$$W = (QX_1, PX_1, QX_2, Z_1)$$

An interesting aspect of this approach is that it makes clear that X_1 plays two roles. (i) estimation of β, and (ii) instrumental variable for Z_2.

This formulation also clarifies the conditions under which it is possible to estimate (identify) both β and γ. Clearly $[QX_1 \vdots QX_2 \vdots Z_1]$ all serve as successful "instruments for themselves". So the question reduces to: are there available IV's for Z_2, the endogenous time invariant variables? This is easily seen to be answered by comparing the number of columns of PX_1 to the number of columns of Z_2. There need to be at least as many columns of PX_1 as the number of columns of Z_2.

Estimating σ_α^2 and σ_u^2. Finally we should address the question of estimating the variances in the matrix Ω. I have two suggestions on this. The first approach relies heavily on the unbalanced nature of the panel, so it is applicable in the problem set, but couldn't be used for the balanced case discussed in the prior discussion of this lecture.

- In the standard balanced panel setting that we have been discussing above, the usual scheme for estimating σ_α^2 and σ_u^2 is as follows:
  - In the within regression, we have residuals, $\hat{u}_i^2$ and compute:

$$\hat{\sigma}_u^2 = (n(T-1))^{-1} \sum_i \sum_t \hat{u}_i^2.$$

  - In the between regression, we have residuals, $\hat{\epsilon}_i = \bar{y}_i - \bar{x}_i\hat{\beta} - z_i\hat{\gamma}$, and compute,

$$\hat{\sigma}_\epsilon^2 = n^{-1} \sum_i \hat{\epsilon}_i^2.$$

  - Since $E\epsilon_i^2 = E(\alpha_i + \bar{u}_i)^2 = \sigma_\alpha^2 + T^{-1}\sigma_u^2$, we can estimate σ_α^2 by,

$$\hat{\sigma}_\alpha^2 = \hat{\sigma}_\epsilon^2 - T^{-1}\hat{\sigma}_u^2.$$

An obvious potential flaw in this strategy is that there is no guarantee that $\hat{\sigma}_\alpha^2 > 0$. Indeed, even in simulations from idealized forms of the model, one regularly encounters negative estimates. Bayesian methods that impose priors supported on $\mathcal{R}_+$ are one way to deal with this problem.

In balanced settings one can also expand the scope of the estimated covariance matrix to acommodate heteroscedasticity and various forms of dependence. This is discussed, for example, in Wooldridge.

- In the unbalanced approach some other methods can be considered. The foregoing approach is problematic since we would like to weight the contributions to σ_u^2 differently for different individuals with different T_i's. Here is one possible option:
  - *Between approach.* In the B-data we have

$$\bar{\varepsilon}_i = \alpha_i + T_i^{-1} \sum_{t=1}^{T_i} u_{it}$$

so

$$V(\bar{\varepsilon}_i) = \sigma_\alpha^2 + T_i^{-1} \sigma_u^2$$

so we have a simple model for heteroscedasticity in this equation, and we can *estimate* by fitting the model

$$\bar{\varepsilon}_i^2 = \sigma_\alpha^2 + \sigma_u^2 (1/T_i)$$

to the squared residuals from the between model.

  - Using the within data as a check of this, we have,

$$\tilde{u}_{it} = u_{it} - \bar{u}_{it}$$

so

$$V(\tilde{u}_{it}) \approx \sigma_u^2$$

and we can then compare $\hat{\sigma}_u^2$ with what we get in the first approach based on the between data.

Any of the standard procedures for estimating these variance components, and it should be noted that procedure 1. is not standard, has the unfortunate possibility that we obtain negative estimates of one of the variances. Again Bayesian methods that impose prior information that both variances should be positive can avoid this at some (possibly considerable) additional computational burden. Sometimes when negative variance components appear in these models, one can blame it on poorly specified models, and using such circumstances as an excuse to rethink the model specification is often a worthwhile step.

References

Hausman J. and W.E. Taylor (1981). Panel Data and Unobservable Individual Effects, *Econometrica*, pp. 1377-98.