Econometrics Lecture 6: Panel Data Analysis

R. G. Pierse

1 Introduction

In this lecture we look at the econometrics of panel data. Panel data refers to the *pooling* of observations on a cross-section of households, firms or countries over several time periods. Panel data is increasingly becoming available, mainly through large surveys, repeated over time. One of the earliest examples is the Panel Study of Income Dynamics which started in 1968 at the University of Michigan. This survey collects annual data on over 5000 variables on about 4800 families in the USA and was designed to investigate the causes of poverty.

Panel data has several advantages. Panels typically have a very large number of cross-sectional observations and so provide large samples for the econometrician to work with. Furthermore, the observations are often at the level of the economic decision making agent (household or firm) and so avoid the problems of *aggregation* implicit in macroeconomic time series data. Panels can be used to look at issues that can not be addressed using pure cross-section or time series data.

On the other hand, there are problems associated with data panels. The panel needs to designed carefully to get representative coverage of the population being studied. If not then there is a problem of selectivity, where the sample is censored because some group in the population is not being observed. There is also a related problem of attrition where observations may drop out of the sample because individuals die, households move or emigrate, firms go bankrupt. This may bias the sample since, for example, only the more successful firms will be observed over time in a panel. Another problem stems from the fact that most panels only have a small number of time observations so that dynamic effects may only be poorly measured.

For a readable introduction to the econometrics of panel data, see Baltagi (1995).

2 Fixed and Random Effects Models

The panel data model can be written as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}, \quad i = 1, \cdots, n \quad t = 1, \cdots, T$$

where y_{it} is the observation on the dependent variable y for the ith cross-sectional unit in the tth period, $\mathbf{x}_{it}$ is a $1 \times k$ vector of observations on k explanatory variables for the ith individual in the tth period, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters.

u_{it} is a disturbance term and we assume that

$$u_{it} = \mu_i + v_{it}$$

so that the error contains an unobservable individual specific effect μ_i and a remainder disturbance v_{it}. μ_i captures characteristics of the individual i that are not picked up by the explanatory variables $\mathbf{x}_{it}$ but which are assumed to be time invariant.

Stacking the observations first by time and then by individual, the model can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{2.1}$$

where $\mathbf{y}$ is an $nT \times 1$ vector, $\mathbf{X}$ is an $nT \times k$ matrix and $\mathbf{u}$ is an $nT \times 1$ vector defined by

$$\mathbf{u} = (\mathbf{I}_n \otimes \boldsymbol{\iota}_T)\boldsymbol{\mu} + \mathbf{v} \tag{2.2}$$

where $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones, $\boldsymbol{\mu}$ is an $n \times 1$ vector of individual specific disturbances and $\mathbf{v}$ is an $nT \times 1$ vector of remainder disturbances.

Two alternative models result from different assumptions about the indiviual specific effects $\boldsymbol{\mu}$.

2.1 The Fixed Effects model

The fixed effects model asumes that the individual specific effects $\boldsymbol{\mu}$ are *fixed* (non-stochastic) parameters to be estimated and that the remaining disturbance component is independently and identically distributed with

$$E(\mathbf{v}) = \mathbf{0}, \quad \mathrm{var}(\mathbf{v}) = \sigma_v^2\mathbf{I}_{nT} \ .$$

On these assumptions, the model (2.1) and (2.2) can be written as

$$\mathbf{y} = \mathbf{D}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{v} \tag{2.3}$$

where $\mathbf{D} = (\mathbf{I}_n \otimes \boldsymbol{\iota}_T)$ is an $nT \times n$ matrix of *dummy variables*. This model can be estimated by *OLS*.

Note that, if $\mathbf{X}$ contains an intercept $\mathbf{c}$, then the set of variables $\mathbf{D}:\mathbf{c}$ will be perfectly collinear. This is the familiar problem of a redundant dummy variable and can be solved either by dropping one of the columns of $\mathbf{D}$ and the associated element of $\boldsymbol{\mu}$ or by imposing the restriction that $\boldsymbol{\iota}_n'\boldsymbol{\mu}=0$.

Often we are only interested in estimating the parameters $\boldsymbol{\beta}$. From the familiar formula for a partitioned regression estimator, we have

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbf{y} \tag{2.4}$$

where $\mathbf{M}$ is the $nT\times nT$ *idempotent* matrix defined by

$$\begin{aligned}
\mathbf{M} &= \mathbf{I}_{nT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' \\
& \quad \mathbf{I}_{nT} - (\mathbf{I}_n\otimes\boldsymbol{\iota}_T)(\mathbf{I}_n\otimes(\boldsymbol{\iota}_T'\boldsymbol{\iota}_T)^{-1})(\mathbf{I}_n\otimes\boldsymbol{\iota}_T') \\
&= \mathbf{I}_{nT} - (\mathbf{I}_n\otimes\frac{\boldsymbol{\iota}_T\boldsymbol{\iota}_T'}{\boldsymbol{\iota}_T'\boldsymbol{\iota}_T}) = \mathbf{I}_{nT} - (\mathbf{I}_n\otimes\frac{1}{T}\boldsymbol{\iota}_T\boldsymbol{\iota}_T')
\end{aligned}$$

and the final equality comes from noting that $\boldsymbol{\iota}_T'\boldsymbol{\iota}_T=T$. Note also that $\boldsymbol{\iota}_T\boldsymbol{\iota}_T'$ is a $T\times T$ matrix of ones.

The matrix $\mathbf{M}$ has the effect of transforming the data by subtracting the individual mean. Thus $\widetilde{\mathbf{y}}=\mathbf{M}\mathbf{y}$ has typical element defined by $\widetilde{y}_{it}=y_{it}-\overline{y}_i$ where $\overline{y}_i=\frac{1}{T}\sum_t y_{it}$. Similarly, $\widetilde{\mathbf{X}}=\mathbf{M}\mathbf{X}$ has typical element defined by $\widetilde{x}_{it,j}=x_{it,j}-\overline{x}_{i,j}$ where $\overline{x}_{i,j}=\frac{1}{T}\sum_t x_{it,j}$. Then (2.4) can be rewritten as

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{y}} \tag{2.5}$$

Equation (2.5) is sometimes known as the *within-groups estimator*.

2.2 The Random Effects Model

The fixed effects assumption leads the inclusion of n dummy variables in the model. When n is large, this results in a considerable loss of degrees of freedom in the estimation of the parameters of interest, $\boldsymbol{\beta}$. An alternative approach is to assume that $\boldsymbol{\mu}$ is random with

$$\mathrm{E}(\boldsymbol{\mu})=\mathbf{0},\quad \mathrm{var}(\boldsymbol{\mu})=\sigma_\mu^2\mathbf{I}_n.$$

As in the fixed effects model, we continue to assume that

$$\mathrm{E}(\mathbf{v})=\mathbf{0},\quad \mathrm{var}(\mathbf{v})=\sigma_v^2\mathbf{I}_{nT}$$

and now make the additional assumption that the two error components $\boldsymbol{\mu}$ and $\mathbf{v}$ are *independent* of each other.

On these assumptions, the combined error term (2.2) has mean zero and variance given by

$$\begin{aligned}\text{var}(\mathbf{u}) &= \mathrm{E}(\mathbf{uu}') = (\mathbf{I}_n \otimes \boldsymbol{\iota}_T)\,\mathrm{E}(\boldsymbol{\mu}\boldsymbol{\mu}')(\mathbf{I}_n \otimes \boldsymbol{\iota}_T') + \mathrm{E}(\mathbf{vv}') \\ &= \sigma_\mu^2(\mathbf{I}_n \otimes \boldsymbol{\iota}_T\boldsymbol{\iota}_T') + \sigma_v^2\mathbf{I}_{nT} = \boldsymbol{\Omega}.\end{aligned}$$

The covariance matrix $\boldsymbol{\Omega}$ is homoscedastic but exhibits serial correlation of the Tth order with

$$\text{cov}(u_{it}u_{js}) = \begin{cases} \sigma_\mu^2 + \sigma_v^2, & i = j, \quad t = s \\ \sigma_\mu^2, & i = j, \quad t \neq s \\ 0 & \text{otherwise.} \end{cases}$$

On the assumption that the regressors $\mathbf{X}$ are *fixed in repeated samples*, *OLS* estimates will be unbiased but inefficient.

The random effects model can be estimated efficiently using *GLS* with the estimator given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}\ . \tag{2.6}$$

This is not a feasible estimator however since the variances σ_μ^2 and σ_v^2 are unknown. Feasible estimates can be based on any consistent estimates of σ_μ^2 and σ_v^2. A consistent estimate of σ_v^2 can be obtained from

$$\widehat{\sigma}_v^2 = \frac{\widetilde{\mathbf{e}}'\widetilde{\mathbf{e}}}{nT - n - k}$$

where $\widetilde{\mathbf{e}}$ are the residuals from the *within-groups estimator* (2.5). A consistent estimator of σ_μ^2 can be obtained from the residuals from the *between-groups estimator*

$$\overline{y}_i = \overline{\mathbf{x}}_i\boldsymbol{\beta} + \overline{u}_i, \quad i = 1, \cdots, n \tag{2.7}$$

leading to the estimator

$$\widehat{\sigma}_1^2 = \frac{\overline{\mathbf{e}}'\overline{\mathbf{e}}}{n - k}$$

where $\overline{\mathbf{e}}$ are the residuals from (2.7) and

$$\widehat{\sigma}_\mu^2 = \frac{1}{T}(\widehat{\sigma}_1^2 - \widehat{\sigma}_v^2)\ . \tag{2.8}$$

Note that in principle from (2.8) it is possible that $\widehat{\sigma}_v^2 > \widehat{\sigma}_1^2$ so that $\widehat{\sigma}_\mu^2 < 0$.

The feasible *GLS* estimator is consistent and asymptotically efficient when either $n \to \infty$ or $T \to \infty$. It is possible to view the *GLS* estimator (2.6) as a linear combination of the *within-groups* and *between-groups* estimators given by

$$\widetilde{\boldsymbol{\beta}} = \mathbf{W}_1\widehat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{W}_1)\overline{\boldsymbol{\beta}}$$

where $\mathbf{W}_1$ is a matrix of weights. As $\sigma_v^2/\sigma_1^2 \to 0$, or as $T \to \infty$, then $\mathbf{W}_1 \to \mathbf{I}$ and $\widetilde{\boldsymbol{\beta}}$ tends to the *within-groups estimator*. Conversely as $\sigma_v^2/\sigma_1^2 \to \infty$, $\mathbf{W}_1 \to \mathbf{0}$ and $\widetilde{\boldsymbol{\beta}}$ tends to the *between-groups estimator*.

2.3 Fixed versus random effects

How can we decide whether to use the fixed or random effects model? Mundlak (1978) suggests an interpretation of the models which leads to an answer to this question. He suggests that in *both* models we should view the effects μ_i as random. However, in the fixed effects model, estimation is done *conditional* on the realised μ_i in the sample. The random effects model estimates the model unconditionally but requires the assumption that the effects μ_i are uncorrelated with the regressors $\mathbf{X}$. When this assumption is valid, then the random effects model uses more information which makes it a more efficient estimator. However, if the assumption of no correlation between μ_i and $\mathbf{X}$ is violated, then the random effects model leads to inconsistent estimates, whereas the fixed effects model is still consistent. Thus if there is uncertainty about whether the effects may be correlated with the regressors, then the fixed effects model may be a safer choice. A test for the validity of the assumption of orthogonality of regressors and errors has been developed by Hausman (1978) and is discussed in the next section.

3 Hypothesis testing in Panel Data Models

3.1 Testing for poolability

One natural question that arises in panel data is whether it is appropriate to pool. This amounts to testing the panel data model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}, \quad i = 1, \cdots, n \quad t = 1, \cdots, T \tag{3.1}$$

against the more general model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta}_i + u_{it}, \quad i = 1, \cdots, n \quad t = 1, \cdots, T \tag{3.2}$$

where the $\boldsymbol{\beta}$ parameters are allowed to differ between individuals. The null hypothesis that pooling is justified is given by

$$H_0 : \boldsymbol{\beta}_i = \boldsymbol{\beta}, \quad i = 1, \cdots, n$$

which imposes $(n-1)k$ restrictions on (3.2).

On the assumption that

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}_{nT})$$

this hypothesis can be tested by the F-statistic

$$\frac{SSR_r - SSR_u}{SSR_u}\frac{n(T-k)}{(n-1)k} \sim F((n-1)k, n(T-k))$$

where SSR_u is the sum of squared residuals from OLS estimation of the unrestricted model (3.2) and SSR_r is the sum of squared residuals from the restricted model (3.1).

3.2 Testing the Random Effects model

A key assumption of the random effects model is that the random effects $\boldsymbol{\mu}$ are not correlated with the regressors $\mathbf{X}$. If this assumption is violated then $E(u_i|\mathbf{X}) \neq 0$ and the GLS estimator is *biased* and *inconsistent*. It is very important therefore to test for the validity of this assumption. Hausman (1978) proposes a *specification test* based on the difference between the *within-groups* and the *between-groups* estimators. The principle of the test is that if the assumption of orthogonality of regressors and errors is violated, then the *within-groups* estimator remains consistent but the *between-groups* estimator is inconsistent.

The test statistic is given by

$$(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}})'\mathbf{V}(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) \sim_a \chi^2_k$$

which is asymptotically distributed as a chi-squared statistic with k degrees of freedom on the null hypothesis that random effects and regressors are uncorrelated. The covariance matrix $\mathbf{V}$ is defined by

$$\mathbf{V} = \mathrm{var}(\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}) = \sigma^2_v(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1} + \sigma^2_1(\mathbf{X}'(\mathbf{I}-\mathbf{M})\mathbf{X})^{-1}.$$

4 Dynamic Panel Data Models

Many economic relationships are dynamic in nature. The dynamic panel data model allows for this by including a lagged dependent variable.

The dynamic panel data model can be written as

$$y_{it} = y_{it-1}\alpha + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}, \quad i = 1, \cdots, n \quad t = 1, \cdots, T \tag{4.1}$$

where α is a scalar coefficient on the lagged dependent variable y_{it-1} and, as before, we assume

$$u_{it} = \mu_i + v_{it}\ .$$

Introducing a lagged dependent variable causes problems in the panel data model. This is because μ_i is correlated with y_{it} and hence with y_{it-1} so that the regressors and error term are correlated. In the random effects model, the error term u_{it} is autocorrelated so that *OLS* estimation of (4.1) will be biased and inconsistent.

In the fixed effects model, Nickell (1981) showed that the *within-groups estimator* leads to biased estimates since the transformed lagged dependent variable $\widetilde{y}_{-1}$ will be correlated with the transformed error $\widetilde{v}$. This bias disappears as $T \to \infty$ but in panel data T is typically small so that biases will be significant.

One alternative approach is to transform the model by first differencing. This results in the transformed model

$$\Delta y_{it} = \Delta y_{it-1}\alpha + \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \quad i = 1, \cdots, n \quad t = 2, \cdots, T \tag{4.2}$$

and

$$\Delta u_{it} = \Delta v_{it}$$

that eliminates the individual effects μ_i but introduces serial correlation (a unit root moving average error) in the disturbances.

The model (4.2) can be estimated by instrumental variables using instruments for Δy_{it-1}. Possible instruments are Δy_{it-2} or y_{it-2} which will not be correlated with Δv_{it} so long as v_{it} are not serially correlated. Arellano and Bond (1991) argue that additional instruments can be found by taking account of the orthogonality conditions in the model. They suggest using the instrument set $\{y_{i1}, y_{i2}, \cdots, y_{it-2}\}$ so that the number of instruments increases with t.

References

[1] Arellano, M. and S. Bond (1991), Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies*, 58, 277–297.

[2] Baltagi, B.H. (1995), *Econometric Analysis of Panel Data*, Wiley, Chichester.

[3] Hausman, J.A. (1978), 'Specification tests in econometrics', *Econometrica*, 46, 1251–1271.

[4] Mundlak, Y. (1978), 'On the pooling of time series and cross section data', *Econometrica*, 46, 69–85.

[5] Nickell, S. (1981), 'Biases in dynamic models with fixed effects', *Econometrica*, 49, 1417–1426.