

# 1 Research Methods 2: Panel Data

When we have observations on individual units over time, for example following the same individuals and interviewing them for a number of years, then we can compare the same individuals in different circumstance. We can basically treat them as their own control

Also have 'data fields' where we collect together data over a number of years for countries or regions

Balanced panels-same no. of time series observations and unbalanced panels-different series length available

Different types of panels

- Same families/individuals followed by survey-often better if long period between them
- Create panel data from cross sections -usually by aggregation
- Use repeated cross sections to create panel data on birth cohorts

Main advantage of panel is that can deal with unobserved heterogeneity

Consider regression:

$$y_{it} = \alpha + \beta x_{it} + \theta_i + \mu_t + \varepsilon_{it}$$

where  $\mu_t$  is time effects that apply to all individuals and  $\theta_i$  the fixed effects for observation  $i$  that are unobservable/unmeasurable. These capture heterogeneity that would cause inconsistency in OLS cross section regression as they are not uncorrelated with the observed independent variables.

As have more than one observation can remove  $\theta_i$  by taking differences. For example if are two surveys

$$y_{i2} - y_{i1} = (\alpha - \alpha) + \beta(x_{i2} - x_{i1}) + (\theta_i - \theta_i) + (\mu_2 - \mu_1) + (\varepsilon_{i2} - \varepsilon_{i1})$$

or if there are more than two time series observations by taking differences from the mean of the variables over the time period

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + (\mu_t - \bar{\mu}) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

Note that doing this means there are  $n(T-1)$  observations rather than  $nT$  and if  $T=2$  it means you lose half of the observations!

regression equations are free from any correlation between the explanatory variable and the unobserved fixed effects and can be estimated consistently by OLS

fixed effects must be fixed over time and must enter additively and linearly

benefits do come with costs:

reduces number of observations and hence efficiency > Need to be careful not to misinterpret a decrease in efficiency as a change in parameter between difference and undifferenced equations. Eg if coeff positive and sig in cross section and becomes insignificant in FE model. need to check if its sig different from cross section value as well as zero.

Differencing removes all fixed effects including regressors that don't change over period of observation. Can remove attraction of procedure esp in short panels

Measurement error: With measurement error in the regressors the difference estimators may no longer be consistent or any better than OLS

## 1.1 More generally

- Suppose we have a panel of data for groups (e.g. people, countries or regions)  $i = 1, 2, \dots, N$  over time periods  $t = 1, 2, \dots, T$  on a dependent variable  $y_{it}$  and an independent variable  $x_{it}$  and we are interested in measuring the effect of  $x_{it}$  on  $y_{it}$ . say

$$y_{it} = \alpha_i + \beta_i x_{it} + \varepsilon_{it} \quad (1)$$

- where  $E(\varepsilon_{it}) = 0$ ;  $E(\varepsilon_{it}^2) = \sigma_i^2$ ;  $E(\varepsilon_{it}\varepsilon_{jt}) = \sigma_{ij}$ ;  $E(\varepsilon_{it}\varepsilon_{jt-s}) = 0$  for  $s \neq 0$ . Notice that here  $k$  does not include the intercept, whereas above it did.

The panel data estimators for the linear model are all standard, either the application of OLS or GLS.

- There are 3 literatures on this type of problem, distinguished by the relative magnitudes of  $N$  and  $T$  and the assumptions that are made about parameter and variance homogeneity.
1. The large  $T$  small  $N$  literature. This uses time-series asymptotics,  $T$  going to infinity  $N$  fixed. The standard model is the Zellner Seemingly Unrelated Regression Estimator, SURE, which estimates the full model above by GLS allowing for the between group covariances  $E(\varepsilon_{it}\varepsilon_{jt}) = \sigma_{ij}$ . Notice that the between group covariance matrix involves estimating  $N(N+1)/2$  elements, so grows rapidly with  $N$ .

2. The large  $N$  small  $T$  literature. This arises typically with large surveys like the BHPS where the number of time periods is small (5 is quite large) but there may be many thousand cross-section observations.  $T$  is not large enough to estimate a model for each group so strong homogeneity assumptions tend to be imposed on the slope parameters  $\beta_i = \beta$  and also often on the intercept parameters,  $\alpha_i = \alpha$ . Between group covariances,  $\sigma_{ij}$ , are assumed to be zero. The asymptotic properties of the estimators are established by letting  $N$  go to infinity,  $T$  fixed. These are usually non-linear models.
3. Large  $N$  large  $T$  literature (sometimes known as panel time-series), where  $T$  is large enough to estimate an equation for each group, but  $N$  is too large to allow for a freely estimated between group covariance matrix. The asymptotics involves letting both  $N$  and  $T$  go to infinity in some way.

## 2 Fixed Effects

The most widely used model in the panel literature is the Fixed Effect (FE) model:

$$y_{it} = \alpha_i + \beta x_{it} + u_{it} \quad (2)$$

$E(u_{it}) = 0; E(u_{it}^2) = \sigma^2$  all  $i; E(u_{it}u_{jt-s}) = 0$  for  $s \neq 0$  and  $i \neq j$ . This restricts the slope coefficients and the variances to be the same across groups, while letting the intercepts differ, and treats all between group covariances as zero. This model is known by a large number of different names, because it was developed independently in many areas. These include:

the “Least Squares Dummy Variable” model (because it can be implemented by running a least squares regression including a dummy (0,1) variable for each group;

the “Within Estimator” since it just uses the within group variation, see below;

the (one way) “Fixed Effects” estimator, in contrast to the two way Fixed Effects and Random Effects estimators discussed below;

the analysis of covariance estimator; and various other names.

Notice that we cannot estimate  $\alpha_i$  consistently ( $N \rightarrow \infty, T$  fixed), since the number of parameters grows with the sample size  $N$ . However we can estimate  $\beta$  consistently.

The total variation in  $y_{it}$  can be decomposed into the within group variation and the between group variation:

$$\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y})^2 = \sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)^2 + T \sum_{i=1}^N (\bar{y}_i - \bar{y})^2$$

where

$$\bar{y} = \sum_{i=1}^N \sum_{t=1}^T y_{it}/NT; \quad \bar{y}_i = \sum_{t=1}^T y_{it}/T;$$

the FE “Within Regression” just uses the within group variation, since the group specific intercepts can be removed by taking deviations from the group mean, allowing (2) to be written:

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + u_{it} \quad (3)$$

the “Between regression” is the cross-section regression using the group means:

$$\bar{y}_i = \alpha + \beta \bar{x}_i + u_i.$$

If the intercepts are all regarded as identical, then one just gets standard OLS on all the data:

$$y_{it} = \alpha + \beta x_{it} + u_{it} \quad (4)$$

or

$$(y_{it} - \bar{y}) = \beta(x_{it} - \bar{x}) + u_{it}. \quad (5)$$

This gives equal weight to the within group and the between group variation.

Two way fixed effect model allows for a separate intercept for every group and every time period:

$$y_{it} = \alpha_i + \alpha_t + \beta x_{it} + u_{it}.$$

Notice that we cannot estimate N+T free intercepts (there would be exact multicollinearity, the dummy variable trap), some restriction is required to identify the parameters and a common one is to express the model as.

$$y_{it} = \alpha + \mu_i + \mu_t + \beta x_{it} + u_{it}$$

subject to  $\sum_{i=1}^N \mu_i = 0$ ,  $\sum_{t=1}^T \mu_t = 0$ . This can be estimated by taking deviations from the year means  $\bar{y}_t$  and  $\bar{x}_t$  and well as the group means,

### 2.0.1

## 2.1 Random Effect Models

The one way fixed effect model involves estimating  $N$  separate  $\alpha_i$  and if  $N$  is large, in the thousands, this involves a lot of parameters and a large loss in efficiency. The alternative is the “Random Effects” model which treats the  $\mu_i$  not as fixed parameters to be estimated, but as random variables,  $E(\mu_i) = 0$ ,  $E(\mu_i^2) = \sigma_\mu^2$ . It is assumed that randomness implies that the  $\mu_i$  are distributed independently of  $u_{it}$  and (the strong assumption) independently of  $x_{it}$ .

With these assumptions we only have to estimate 2 parameters  $\alpha$  and  $\sigma_\mu^2$  not the  $N$   $\alpha_i$ . The model is then:

$$y_{it} = \alpha + \beta x_{it} + (u_{it} + \mu_i)$$

where the parentheses indicate the new error term,  $v_{it} = (u_{it} + \mu_i)$ .  $E(v_{it}) = 0$ ;  $E(v_{it}^2) = \sigma^2 + \sigma_\mu^2$ ;  $E(v_{it}v_{it-s}) = \sigma_\mu^2$ ,  $s \neq 0$ ;  $E(v_{it}v_{jt-i}) = 0$ ,  $i \neq j$ . Thus this error structure introduces a very specific form of serial correlation. Estimation is by Generalised Least Squares,

## 2.2 Testing

If the unrestricted model is, the general model with no between group covariances

$$y_{it} = \alpha_i + \beta_i x_{it} + \varepsilon_{it} \quad (6)$$

and the restricted model is the fixed effect model

$$y_{it} = \alpha_i + \beta x_{it} + u_{it}. \quad (7)$$

This appears to be merely involve testing equality of the slope coefficients, i.e. the  $k(N - 1)$  restrictions  $\beta_i = \beta$ ,  $i = 1, 2, \dots, N$ . The standard F test (Chow test) for this is:

$$\frac{(\sum \sum \hat{u}_{it}^2 - \sum \sum \hat{\varepsilon}_{it}^2)/N(k - 1)}{\sum \sum \hat{\varepsilon}_{it}^2/(NT - N(k + 1))} \sim F[k(N - 1), (NT - N(k + 1))].$$

The difficulty is that this test will only be correct if the variances are the same across groups:  $\sigma_i^2 = \sigma$  for all  $i$ .

One alternative is to use Likelihood Ratio Tests which can be calculated from the same two sets of least squares regressions. If both coefficients and variances differ, the maximised log likelihood is, apart from a constant, the sum of the log likelihoods for the individual equations:

$$L_1 = -\frac{T}{2} \sum_{i=1}^N \ln \hat{\sigma}_i^2; \quad \hat{\sigma}_i^2 = \sum_{t=1}^T \hat{\varepsilon}_{it}^2.$$

If the coefficients differ, but the variances are the same, the maximised log likelihood is:

$$L_2 = -\frac{NT}{2} \ln \hat{\sigma}^2; \quad \hat{\sigma}^2 = \sum_{i=1}^N \sum_{t=1}^T \hat{\varepsilon}_{it}^2.$$

If both the slope coefficients and the variances are the same, the maximised log likelihood is:

$$L_3 = -\frac{NT}{2} \ln \tilde{\sigma}^2; \quad \tilde{\sigma}^2 = \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2.$$

There is a fourth case, equal coefficients and different variances, discussed above.

The test for equality of variances is then just  $2(L_1 - L_2) \sim \chi^2(N - 1)$ . The test for equality of both coefficients and variances is just  $2(L_1 - L_3) \sim \chi^2(N - 1)(k + 1)$ . The LR equivalent of the F test above (equality of coefficients conditional on equality of variances) is  $2(L_2 - L_3) \sim \chi^2(k[N - 1])$ . Exactly the same sort of procedure can be used for testing equality of intercepts. If  $N$  was small we could start from the more general SURE model, which allowed for between group covariances.

The Likelihood ratio approach does not work with the Random Effects model, since it is a GLS rather than ML estimator. The usual approach is

to test between OLS and RE, by using a standard LM test for heteroskedasticity, since the variances will differ between groups if the RE model is appropriate. Of course, you may get heteroskedasticity for other reasons than random effects.

To test between RE and FE a Hausman test is used. This uses a test statistic which is the ratio of the squared differences between the fixed and random model results and the difference between their variances, which is distributed  $\chi^2(k)$ . If this is large (the difference between the estimates is significant) you reject the null hypothesis that the RE model is appropriate against the alternative that the FE model is appropriate.

### **2.2.1**

### 3 Dynamics.

- Consider a dynamic version of the fixed effect model

$$y_{it} = \alpha_i + \beta x_{it} + \lambda y_{i,t-1} + u_{it} \quad (8)$$

- the usual estimator is inconsistent ( $N \rightarrow \infty, T$  fixed), because of the usual problem of the downward bias of the lagged dependent variable because of dependence on initial conditions, though the bias declines with  $T$ .
- There are various instrumental variable estimators which are consistent, which remove the  $\alpha_i$  by differencing rather than by taking deviations from the group means.
- If you difference, you get

$$\Delta y_{it} = \beta \Delta x_{it} + \lambda \Delta y_{i,t-1} + \Delta u_{it} \quad (9)$$

- but  $\Delta u_{it} = u_{it} - u_{i,t-1}$  is clearly correlated with  $\Delta y_{i,t-1} = y_{it} - y_{i,t-1}$  since  $u_{i,t-1}$  determines  $y_{i,t-1}$ .
- However, you can use  $y_{i,t-2}$  and earlier as instruments.
- Suppose the coefficients differ:

$$y_{it} = \alpha_i + \beta_i x_{it} + \lambda_i y_{i,t-1} + u_{it} \quad (10)$$

- and this is ignored, then the equation is:

$$y_{it} = \alpha_i + \beta x_{it} + \lambda y_{i,t-1} + [(\beta_i - \beta)x_{it} + (\lambda_i - \lambda)y_{i,t-1} + u_{it}] \quad (11)$$

- where  $v_{it} = [(\beta_i - \beta)x_{it} + (\lambda_i - \lambda)y_{i,t-1} + u_{it}]$  is the new error term.
- This error term is going to be serially correlated and correlated with the lagged dependent variable, so the estimates will be inconsistent even for large  $T$ .
- This heterogeneity bias, biases  $\hat{\lambda}$  upwards (i.e. in the opposite direction to the initial condition bias) and  $\hat{\beta}$  downwards towards zero.
- If  $x_{it}$  is I(1) serial correlation coefficient of unity and  $\lambda_i = 0$ ,  $P \lim(\hat{\beta}) = 0$ ,  $P \lim(\hat{\lambda}) = 1$ ,  $T \rightarrow \infty$ . The estimates are wrong.
- The bias on the long-run coefficients  $\beta/(1 - \lambda)$  is smaller because the two biases cancel out to some extent.

### 3.0.2

### 3.0.3 Alternatives to Pooling

- If  $T$  is large enough to estimate an equation for each group, rather than imposing homogeneity, which can have adverse consequences (particularly in dynamic models), it may be better to use a weighted average of the individual coefficients of the form
- There are a number of such estimators available, differing in the choice of weights.
- One the Swamy Random Coefficients Models, weights inversely to the adjusted variances of the  $\hat{\beta}$

### Instrumental Variables

- There are a number of methods that deal with the dynamics using instrumental variable approaches. Arrelano and Bond in particular.
- This was developed for small T methods and is available in STATA, but not Limdep
- We will not consider here:
- Deatons suggests we should not expect too much of such models

## 3.1 Concluding remarks

- Having panel data can provide benefits for analysis but it depends upon the questions you want to ask, the type of data you have and the size of the data (N and T)
- There are useful procedures readily available in the packages: Fixed and random effects have advantages over pooling and cross section analysis
- We can deal with dynamic models though there are problems
- But should compare the results from the various methods and be clear about question we are trying to answer
- Panel data models are a growing area of econometric analysis, particularly dynamic models. many of the more complicated models are not readily available, but can go a long way with the simpler ones. Eg dynamic fixed effects models are informative with large T...
- Having knowledge of this area is extremely valuable