PANEL DATA

· If we have individual cross sectional data in for several years we have a panel

· Panel data: vary across units, and for the same unit across time => two dimensions (sources) of variation

· Denote y_{it} a variable for individual i at time t and x_{it} some explanatory variable

· Assume n individuals and T periods=> nxT observations

· Panel data can have more than two dimensions. For instance:

- n multinational firms
- Over T periods
- For C countries (e.g. sales of same firm but in a different country)

· Time variation is not necessary to characterize a panel: what matters is that variation has more than just one dimension

- n multinational firms, for C countries is still a panel

· Panel data often arise in matched data and there are multiple matches:

- A sample of firms that borrow from several banks
- A sample of workers that serve different firms
- A sample of teachers that teach different classes

· Panels can be:

- Balanced: units are followed for an equal number of periods (firms borrow from same number of banks)

- Unbalanced: units are followed for different number of periods

- Continuous: where there is no entry of new units. Start with n units at $T = 0$ and over time units are not replaced if they exit (exit is called attrition)

- Rotating: units are followed for a fixed number of periods (minimum two), then units leave and are replaced by a new sample

Types of panel data

Unit ID (firms, household, country, region etc.)	Year (or code of unit over which ID varies)	x_{it}
001	1990	100
001	1991	120
001 T	1992	90
001	1993	77
002	1990	1250
002	1991	1370
002 T	1992	2220
002	1993	2001
....		
NxT		

Balanced

Unit ID (firms, household, country, region etc.)	Year (or code of unit over which ID varies)	x_{it}
001	1990	100
001 T_1	1991	120
001	1992	90
001	1993	77
002	1990	1250
002 T_2	1991	1370
003	1991	37000
003 T_3	1992	38560
003..	1993	41000

Unbalanced

A panel data set contains n entities or subjects (e.g., firms and states), each of which includes T observations measured at 1 through t time period. Thus, the total number of observations is nT. Ideally, panel data are measured at regular time intervals (e.g., year, quarter, and month).
Otherwise, panel data should be analyzed with caution.

A *short panel data* set has many entities but few time periods (small *T*), while a *long panel* has many time periods (large T) but few entities (Cameron and Trivedi 2009: 230).

Why panel data are interesting

· Allow to control for unobserved heterogeneity: if unobserved heterogeneity remains fixed it can be handled

· May offer statistical advantages: for instance help solve multicollinearity (as extra dimension of variation is added)

· May help address issues of dynamics if the panel has time variation

Main disadvantage of panel data: attrition

· People initially in the panel may disappearas time elapses. Problem: those who leave may not do so randomly (in general they will not) => can give rise to selection.

· Example 1: dynamics of poverty may seem to improve if the poor are more likely to leave the survey (e.g. more likely to be fired and need tomove to find another job)

· Example 2: in a panel of firms profitability mayseem to improve, but effect may be due to survival of the fittest, while inefficient firms fail and exit the economy and the sample

Pooled Data

· One could simply ignore panel nature of data and estimate:

$$y_{it}=\beta' x_{it}+\varepsilon_{it}$$

· This will be consistent if

$$\text{plim } (X' \varepsilon/N)=0$$

· But computed standard errors will only be consistent if errors uncorrelated across observations

· This is unlikely:
- Correlation between errors of different individuals in same time period (aggregate shocks) => deal with time dummies

- Correlation between errors of same individual in different time periods: may arise if there is some unobserved heterogeneity (e.g. unobserved ability that is a fairly stable characteristic of the

individual).

A More Interesting Model

· Here a single explanatory variable (but can think of x as a vector)

· f_i is a variable specific to the individual i and time invariant : i.e. an individual component (for instance ability in the returns to education model)

· Panel data models differ in the interpretation of f_i

Three Models

· **Random Effects Model**

- Treats f_i as part of error term (in addition to ε_{it})
- Consistency does require no correlation between f_i and x_{it} : hence this model assumes $E(f_i | x_{it}) = 0$

· Fixed Effects Model

- Treats f_i as parameters to be estimated (like γ)
- Consistency does not require anything about correlation with x_{it} : this model assumes $E(f_i | x_{it}) \neq 0$

· Between-Groups Model

- Runs regression on averages for each individual (i.e. take time averages)

When is FE useful

Example: want to estimate a production function

$$y_{it} = \alpha + \gamma l_{it} + \delta m_i + \varepsilon_{it}$$

$y = \text{output}, \quad l = \text{labor input}, \quad m = \text{managerial ability (time invariant)}$

· Ability is not observed. If only cross sectional data where available one would estimate

$$y_i = \alpha + \gamma l_i + v_i$$

· Ability ends up in the error term and if it is correlated with l_i OLS estimates of production function are inconsistent

· Panel data solve this problem as the fixed effect f_i picks up unobserved ability in

$$y_{it} = \alpha + \gamma l_{it} + f_i + v_{it}$$

· But you need an extra source of variation. What is this? Variation over time!

Back to FE

· Recall main model

$$y_{it} = \alpha + \gamma x_{it} + f_i + \varepsilon_{it}$$

· The key assumption here is that E(εit|xit)=0 but E(fi|xit)≠0

· There is truly unobserved heterogeneity and these omitted fixed factors may be correlated with the explanatory variables

· In this context *fi* can be interpreted as just one of the various parameters to be estimated: since there are *n* individuals and f_i is individual specific, we have to estimate *n* additional parameters

· In other words we estimate a constant term that is individual specific

· If a constant appears in the regression we identify *n*-1 fixed effects. If we omit the constant we identify *n* fixed effects

Estimating the Fixed Effects Model

The LSDV model

· We can estimate the model by including a separate dummy for each individual: this is the **Least Squares Dummy Variable Model** and estimate

$$y_{it} = \alpha + \gamma x_{it} + f_1 d_1 + f_2 d_2 + .. f_{N-1} d_{N-1i} + \varepsilon_{it}$$

Issues and properties of LSDV procedure

· Property: LSDV gives consistent estimates as εit uncorrelated with X (by assumption!)

· If feasible, it is a very simple way to estimate the parameters of interest

· Problem: may be computationally unfeasible if *n* is very large

- Example:dataset has 30,000 firms per year, hard to invert a 30,000 x 30,000 matrix

· Need to find a simpler procedure

· The solution is to "de-mean" the data and

estimate the de-meaned model

Compute for each i (and each explanatory variable if many x)

$$\bar{y}_i = \frac{1}{T}\sum_1^T y_{it}; \quad \bar{x}_i = \frac{1}{T}\sum_1^T x_{it};$$

$$\tilde{y}_{it} = y_{it} - \bar{y}_i; \quad \tilde{x}_{it} = x_{it} - \bar{x}_i$$

Run OLS on

$$\tilde{y}_{it} = \gamma\, \tilde{x}_{it} + \varepsilon_{it}$$

the estimate of γ is the same as in the LSDV model.

This model is easy to compute (need just to compute as many means as individuals)

Why does de-meaning work?

$$y_{it} = \gamma x_{it} + f_i + \varepsilon_{it}$$

$$\bar{y}_i = \gamma \bar{x}_i + f_i + \bar{\varepsilon}_i$$

$$\tilde{y}_{it} = \gamma \tilde{x}_{it} + \tilde{\varepsilon}_{it}$$

Retrieving the fixed effects

One can show that in this case the estimated fixed effects are equal to

$\hat{f}_i = \bar{y}_i - \hat{\gamma}\bar{x}_i$

(from $\bar{y}_i = \gamma\bar{x} + f_i + \bar{\varepsilon} \Rightarrow \bar{y}_i = \hat{\gamma}\bar{x} + f_i + \bar{u} \Rightarrow \bar{y}_i = \hat{\gamma}\bar{x} + \hat{f}_i$

$\Rightarrow \hat{f}_i = \bar{y}_i - \hat{\gamma}\bar{x}$)

This has a nice interpretation: the fixed effect captures all the gap between the observed mean value of y ($\bar{y}_i$) and the predicted mean value of y on the basis of the time variation in x

=> any other variation that can explain $\bar{y}_i$ that is not time varying is reflected in the fixed effect

Features of fixed effect estimator

· Only uses variation **within** individuals - that is why called 'within-group' estimator

· This variation may be a small part of total (so low precision) and more prone to measurement error (so more attenuation bias)

· Cannot use it to estimate effect of regressors

that are constant for an individual, such as race,
schooling, gender etc. All these effects are lumped together in the fixed effect
· They can be estimated if **their effect** is time
varying (e.g. the effect of schooling not the same
across years)

Random Effects Estimator

- Treats f_i as part of residual which becomes $v_{it} = f_i + \varepsilon_{it}$
- Assume , f_i and ε_{it} are both orthogonal to x_{it} at all lags. This is required for consistency
- Assume also that f_i and ε_{it} are uncorrelated with each other and ε_{it} is serially independent
- Even if the OLS estimate of $y_{it} = \alpha + \gamma x_{it} + f_i + \varepsilon_{it}$ is consistent it is not efficient as the "global" error v_{it} is heteroschedastic

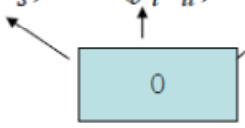
Consider

1) $E(v_{it}^2) = E((f_i + \varepsilon_{it})^2) = E(f_i^2) + E(\varepsilon_{it}^2) + E(f_i \varepsilon_{it})$

$= \sigma_f^2 + \sigma_\varepsilon^2$

2) $E(v_{it} v_{is}) = E((f_i + \varepsilon_{it})(f_i + \varepsilon_{is}) = E(f_i^2) + E(\varepsilon_t \varepsilon_s) + E(f_i \varepsilon_{it}) + E(f_i \varepsilon_{is})$

$= \sigma_f^2$



Which implies that v_{it} for a generic individual i will be heteroschedastic

Namely, for individual i the variance covariance matrix of v_{it} is

$$E(v_i' v_i) = \begin{bmatrix} \sigma_f^2 + \sigma_\varepsilon^2 & \sigma_f^2 & \ldots\ldots\ldots & \sigma_f^2 \\ \sigma_f^2 & \sigma_f^2 + \sigma_\varepsilon^2 & \ldots\ldots\ldots & \sigma_f^2 \\ \sigma_f^2 & \sigma_f^2 & \sigma_f^2 + \sigma_\varepsilon^2 \ldots\ldots & \sigma_f^2 \\ \sigma_f^2 & \sigma_f^2 & \ldots\ldots\ldots & \sigma_f^2 + \sigma_\varepsilon^2 \end{bmatrix} = \Omega_{(T \times T)}$$

and thus for the error term in the model

$$E(v'v) = \begin{bmatrix} \Omega & 0 & 0 \ldots\ldots\ldots & 0 \\ 0 & \Omega & 0 \ldots\ldots\ldots & 0 \\ 0 & 0 & \Omega\ 0 \ldots\ldots & 0 \\ \ldots & \ldots & \ldots\ \ldots\ldots & \ldots \\ 0 & 0 & \ldots\ldots\ldots & \Omega \end{bmatrix} = \Omega \otimes I_{(NT \times NT)}$$ (a block diagonal matrix)

Random Effects Estimator

- To obtain an efficient estimator one relies on the Generalized Least Squares Estimator
- Suppose the variance covariance matrix of $v_{it} = f_i + \varepsilon_{it}$ is known
- One can the use GLS estimator to obtain an efficient estimator
- Idea: transform the heteroschedastic error term into a homoschedastic one and apply OLS to the transformed model. This is the RE estimator

$$\hat{\beta}^{RE} = \left(X' \hat{\Omega}^{-1} X \right)^{-1} X' \hat{\Omega}^{-1} y$$

RE estimator in practice

- Applying GSL to the model with RE model one obtains the RE estimator
- In general however one does not know Σ. I will not describe how to obtain an estimate of Σ (see class notes)
- The RE estimator that uses an estimate of Σ is called feasible GLS (FGLS). The estimator is then

$$b_{RE} = \left(X' \hat{\Sigma}^{-1} X \right)^{-1} X' \hat{\Sigma}^{-1} y$$

Comments

- Assumption about exogeneity of errors is stronger than for FE model – need to assume ε_{it} uncorrelated with whole history of x (called strong exogeneity)
- If exogeneity assumptions are satisfied RE estimate will be more efficient than FE estimator
  - Application of general principle that imposing true restriction on data leads to efficiency gain.

Between-Groups Estimator

· This takes individual means (over time) and estimates the regression by OLS:

$$\overline{y}_i = \alpha + \gamma \overline{x}_i + f_i + \overline{\varepsilon}_i$$

· The error term is now vit = *f*i + εi and need to assume regressors are orthogonal to achieve consistency

· But BE estimator less efficient as it does not exploit variation in regressors for a given individual

· And cannot estimate variables like time trends whose average values do not vary across individuals

· So why would anyone ever use it?

· Main reason: can help reduce measurement error

· Intuition: if time variation is plagued by classical
measurement error, averaging over time reduces
measurement error

· The Fixed Effect estimator is prone to exposure to measurement error.

· Intuition: neglecting between variation eliminates a lot of signal and measurement error bias depends on noise to signal ratio.

· Can also be shown that attenuation bias tends to be larger in FE that RE model => measurement error considerations may affect choice of model to estimate with panel data