ECO 372: Introduction to Econometrics

Spring 2025

Introduction to Panel Data and Traditional Models

Sakib Bin Amin, Ph.D.

Associate Professor in Economics
Director, Accreditation Project Team (APT)

Outline

— Our objectives for this lecture will be to learn:

- ❑ Overview of Panel Data
- ❑ Benefits of Panel Data
- ❑ Limitations of Panel Data
- ❑ History of Panel Data Models
- ❑ Fixed Effects (FE) and Random Effects (RE) Models
- ❑ STATA Example

Overview of Panel Data

- Panel data (also called longitudinal data) refers to data for N different entities observed at T different periods.
- For example, GDP data of 5 South Asian countries from 1980-2024. Here N=5 and T=45, leading to a total NT=5×45=225 observations (each country has 45 observations of GDP).
- When describing cross-sectional data, it was useful to use a subscript to denote the entity. For example , Y_i referred to the variable Y for the for the i^{th} entity.
- When describing panel data, we need some additional notation to keep track of both the entity and the time period. For example, Y_{it} denotes the variable Y for the for the i^{th} entity in the t^{th} period.
- A balanced panel has all its observations; that is, the variables are observed for each entity and each time period.
- A panel that has some missing data for at least one time period for at least one entity is called an unbalanced panel.

Illustration of a Sample Panel Dataset

YEAR (T)	Country (N)	LNCPI	LNGDP	LNINV
1980	BGD	2.278	24.078	21.686
1981	BGD	2.346	24.147	21.969
1982	BGD	2.489	24.169	21.892
1983	BGD	2.830	24.207	21.794
2024	BGD	2.912	24.254	21.861
1980	IND	3.131	24.286	21.984
1981	IND	3.190	24.327	21.982
1982	IND	3.284	24.364	22.048
2024	IND	3.355	24.388	22.154
1980	PAK	3.414	24.416	22.258
1981	PAK	3.473	24.471	22.372
1982	PAK	3.535	24.505	22.378
2024	PAK	3.571	24.558	22.426
1980	SLK	3.601	24.604	22.507
1981	SLK	3.652	24.642	22.550
1982	SLK	3.750	24.692	22.705
2024	SLK	10.774	10.737	11.988
1980	NPL	3.826	24.780	23.077
1981	NPL	3.906	24.831	23.126
1982	NPL	3.850	24.930	23.060
2024	NPL	4.550	20.050	22.600

Benefits of Panel Data

- **Panel data controls for individual heterogeneity.**
  - Panel data suggests that individuals, firms, states or countries are heterogeneous.
  - With Panel data modelling, we can understand the heterogeneity and its impact on choice of variables.
- **Panel data give more informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency.**
  - Time-series models can lead to singnificant multicollinearity if not carefully designed (i.e., due to the use of past values and shocks).
- **Panel data are better able to study the dynamics of adjustment.**
  - Cross-sectional distributions that look relatively stable hide a multitude of changes.
  - For example, in measuring unemployment, cross-sectional data can estimate what proportion of the population is unemployed at a point in time. Repeated cross-sections can show how this proportion changes over time.
  - Only panel data can estimate what proportion of those who are unemployed in one period can remain unemployed in another period.

Benefits of Panel Data

- **Panel data are better able to identify and measure effects that are simply not detectable in pure cross-section or pure time-series data.**
  - For instance, intra-industry labour force participation analysis resulting from wage rigidities. With the use of industry individual effects, panel data models can easily show differences while corss-section and time series models fail to capture such effect.
- **Panel data models allow us to construct and test more complicated behavioral models than purely cross-section or time-series data.**
  - For example, technical efficiency is better studied and modeled with panels
- **Micro panel data gathered on individuals, firms and households may be more accurately measured than similar variables measured at the macro level.**

Limitations of Panel Data

- **Design and data collection problems:**
  - Mainly happens in micro level panel data
  - These include problems of coverage(incomplete account of the population of interest), nonresponse, recall (respondent not remembering correctly), frequency of interviewing, etc.
- **Distortions of measurement errors:**
  - Measurement errors may arise because of faulty responses
  - Misreporting of year to year data
- **Short time-series dimension:**
  - Short time-series may lead to inconsistent results. This happens mainly because the loss of degrees of freedom as a result of lagged or non-lagged variables
- **Cross-section dependence:**
  - Macro panels on countries or regions with long time-series that do not account for cross-country dependence (i.e., inter dependence of variables across the entities) lead to misleading inference.
  - Recent development of panel data modelling can address this issue.

History of Panel Data Models

1940s:

- Development of formal framework for panel data models .

1950-60s:

- Further development of modelling framework.
- Discussion of why OLS will not work.
- Fixed Effects (FE) and Random Effects(RE) Models to address inconsistencies of OLS.

1970-80s:

- The introduction of dynamic panel estimations.
- Issues with dynamic panel estimations when FE model is used.
- Further developments of dynamic panel models with ARDL and DL approaches.
- Development of panel instrumental variables (like GMM and its other variants).

1990s-2000:

- Panel cointegration models.

History of Panel Data Models

2000-2010s:

- Cross-sectional dependence detection and estimation models.
- Heterogeneous slope detection and estimation models.
- Quantile regressions with panel data (e.g., quantile via moment estimator).

2020-till date:

- Theoretical developments on different types of panel estimation models.
- For example, inconsistency reduction in short panels.

Fixed Effects (FE) Model

- Entities have individual characteristics that may or may not influence the outcome and/or predictor variables.
- For example, the business practices of a company may influence its stock price (predictor) or level of spending (outcome).
- Since individual characteristics are not random and may impact the predictor or outcome variables, we need to control for them.
- Otherwise these characteristics will be captured by the error term. As a result, we might face omitted variable bias [i.e., $corr(Y_{it}, \varepsilon_{it})$] and endogeneity bias [i.e., $corr(X_{it}, \varepsilon_{it})$].
- Individual characteristics are often unobservable. To account their effect we generally use dummy variables. Therefore, these effects are time-invariant (constant over time but changes with entities).

Fixed Effects (FE) Model

- The general framework of a FE model is as follows:

$$y_{it} = \alpha_0 + \alpha x_{it} + \beta D_i + \varepsilon_{it}$$

y_{it} =outcome variable

x_{it} = predictor or regressor

D_i= entity specific FE (Dummy: $entity\ 1 = 1\ or\ 0, \ldots, entity\ n = 1\ or\ 0$)

ε_{it}= error term

- Recall, if there are multiple dummies, we can face dummy variable trap. To avoid it, we can use (n-1) dummies. So, if there are 5 FEs, we need to use 4 in the model. The unused FE will act as a benchmark (captured by the constant).
- FE can also be made in terms of time period. For example we can create a dummy for each year in the sample period. For example if we have 10 years of data, we can create 10 dummies as FE. These are known as time fixed effects.
- We can account both entity and time FE in the same model as:

$$y_{it} = \alpha_0 + \alpha x_{it} + \beta D_i + \beta T_t \varepsilon_{it}$$

Fixed Effects (FE) Model

- FE models as discussed sometimes may be difficult to estimate when the number of FEs is very high.
  - Imagine you are creating thousands of dummies, which is a treacherous work in itself.
  - Given the sample size, adding thousands of dummies would certainly lead to reduced degrees of freedom and there will higher risk of multicollinearity.
- As a result, statistical software uses an innovative approach. This is known as the time-demeaned method or partial out method.
- Suppose we have the following model:

$$y_{it} = \alpha_0 + \alpha x_{it} + \beta D_i + \varepsilon_{it}$$

- We sum and take average the of the variables(e.g., $\bar{y}_i = \sum y_{it}/T$). In this way D_i and α_0 will stay the same (because they are constant having value 1).
- Subtract from the main model and that will give a reduced form for estimation.

$$y_{it} - \bar{y}_i = \alpha_0 - \alpha_0 + \alpha(x_{it} - \bar{x}_i) + \beta(D_i - D_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$\tilde{y}_{it} = \alpha \tilde{x}_{it} + \tilde{\epsilon}_{it}$$

Random Effects (RE) Model

- The rationale behind Random Effects (RE) model is that, unlike the FE model, the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model.
- If one has reason to believe that differences across entities have some influence on your dependent variable but are not correlated with the predictors then he(she) should use RE model.
  - Which is however, very difficult job to be done!
- Suppose we have the following model:

$$y_{it} = \alpha_0 + \alpha x_{it} + \beta D_i + \varepsilon_{it}$$

- It will be considered as a RE model if:

$$corr(y_{it}, \varepsilon_{it}) = 0 \;\&\; corr(x_{it}, \varepsilon_{it}) = 0$$

How to Choose between FE and RE?

- Which approach is better for a model we want to estimate?
- We can use the Hausman test to determine whether we should use FE or RE.
- The null hypothesis of this test is: RE is more appropriate. The alternative hypothesis is: that FE is not appropriate.
- If the null hypothesis is rejected, we can confirm that FE needs to be used.

STATA Example

- Suppose we want to estimate the following model

$$lnWage_{it} = \alpha_0 + \beta_1 ln\_age_{it} + \beta_2 ln\_ttl\_exp_{it} + \beta_3 \mu_i + u_i$$

$lnWage_{it}$ =natural log of wage

ln_age_{it} = natural log of age (years)

$ln\_ttl\_exp_{it}$= natural log of total experience (years)

μ_i= FE by employees

STATA Example

```
Fixed-effects (within) regression               Number of obs      =     28,489
Group variable: idcode                          Number of groups   =      4,709

R-sq:                                           Obs per group:
     within  = 0.1292                                         min =          1
     between = 0.2375                                         avg =        6.0
     overall = 0.1660                                         max =         15

                                                F(2,23778)         =    1764.25
corr(u_i, Xb)  = 0.1642                         Prob > F           =     0.0000
```

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_age	.0832587	.0209411	3.98	0.000	.0422128	.1243047
ln_ttl_exp	.1234857	.0050684	24.36	0.000	.1135514	.13342
_cons	1.217351	.0635995	19.14	0.000	1.092692	1.342011
sigma_u	.37944284					
sigma_e	.29886095					
rho	.61714582	(fraction of variance due to u_i)				

```
Random-effects GLS regression                   Number of obs      =     28,489
Group variable: idcode                          Number of groups   =      4,709

R-sq:                                           Obs per group:
     within  = 0.1288                                         min =          1
     between = 0.2388                                         avg =        6.0
     overall = 0.1686                                         max =         15

                                                Wald chi2(2)       =    4733.39
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =     0.0000
```

ln_wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ln_age	.0127519	.0176557	0.72	0.470	-.0218525	.0473564
ln_ttl_exp	.1508213	.0042138	35.79	0.000	.1425625	.1590801
_cons	1.416817	.0545043	25.99	0.000	1.30999	1.523643
sigma_u	.33150116					
sigma_e	.29886095					
rho	.55164164	(fraction of variance due to u_i)				

STATA Example

	Coefficients (b) fixed	(B) random	(b-B) Difference	sqrt(diag(V_b-V_ S.E.
ln_age	.0832587	.0127519	.0705068	.0113247
ln_ttl_exp	.1234857	.1508213	-.0273356	.0028314

b = consistent under Ho and Ha; obtained from x
B = inconsistent under Ha, efficient under Ho; obtained from x

Test: Ho: difference in coefficients not systematic

chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 219.27
Prob>chi2 = 0.0000

The null hypothesis is rejected, we can confirm that FE model is the better choice here.

STATA Example

```
//sample dataset
webuse nlswork
        xtset idcode                    //declaring data as panel data
// create variables
g ln_age=log(age)
g ln_ttl_exp=log(ttl_exp)


//fixed effects model
xtreg ln_w ln_age ln_ttl_exp,fe
        estimates store fixed           //storing fe results for hausman test
//random effects model
xtreg ln_w ln_age ln_ttl_exp,re
        estimates store random          //storing re results for hausman test
//hausman test
hausman fixed random, sigmamore
```