

# Applied Econometrics (MSc.)

## Lecture 8

### Static Linear Panel Data Models

Yonas Alem

Department of Economics  
University of Gothenburg

January 11, 2015

# Introduction

- A panel data set contains repeated observations over the same units (individuals, households, firms, countries) collected over a number of periods
- Can be micro or macro level
- Allows estimation of more complicated and more realistic models than a single cross-section or a single timeseries data
- Some practical limitations:
  - Difficult to assume that different observations of an individual or a household are independent, which results in some complications in non-linear and dynamic models.
  - Very often, panel data suffer from missing observations (attrition).

Table: Panel Data - Example

id	t	x1	x2	x3
1	2000	250	1	45
1	2001	275	1	46
1	2002	322	1	47
2	2000	500	0	29
2	2001	550	0	30
2	2002	600	0	31
3	2000	175	1	36
3	2001	225	1	37
3	2002	305	1	38

## Introduction Contd...

- Consider the regression:

$$y_{it} = \beta_0 + x'_{it}\beta + \varepsilon_{it} \quad (1)$$

- Where  $x_{it}$  is a K-dimensional vector variables, ( $i = 1, \dots, N$ ) & ( $t = 1, \dots, T$ ).
- This model imposes that  $\beta_0$  and  $\beta$  are identical for all individuals and time periods.
- One can use OLS provided the usual assumptions hold.

## Introduction Contd...

- It would however be unrealistic to assume that the  $\varepsilon'_{it}$ s are uncorrelated over time
- Thus, the standard errors from OLS based on i.i.d error terms will be misleading in panel data applications
- Moreover, OLS will be inefficient relative to an estimator that exploits the correlation overtime in  $\varepsilon'_{it}$ s
- A typical panel data model assumes that

$$\varepsilon_{it} = \alpha_i + u_{it} \quad (2)$$

- Where  $u_{it}$  is assumed to be homoscedastic and uncorrelated over time

# Introduction Contd...

- $\alpha_i$  is time invariant and homoscedastic across individuals
- The model specified by (1) and (2) is called an *error components* or *random effects model*.
- We'll see that this estimator (which uses a generalized least square estimator), under some conditions, yields efficient parameter estimates than OLS
- The assumption,  $E(x_{it}\varepsilon_{it}) = 0 \implies x_{it}$  are uncorrelated with the unobservable characteristics in both  $\alpha_i$  and  $u_{it}$
- In many cases however this assumption appears to be restrictive and it is the case that  $E(x_{it}\alpha_i) \neq 0$
- In the case of the wage equation, can you think how this assumption would be unrealistic?

## Introduction Contd...

- In a cross-sectional context, the standard approach to deal with this is to use IV methods.
- With panel data however, it is possible to exploit the particular nature of the data owing to the availability of repeated observations on the same units.
- In a *fixed effects model*, this problem is addressed by including an individual-specific intercept term in the model as follows

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it} \quad (3)$$

- Where  $\alpha_i$ , ( $i = 1, \dots, N$ ) are fixed unknown constants that are estimated along with  $\beta$  and where  $u_{it}$  is typically assumed to be i.i.d over individuals and time.
- $\alpha_i$  is referred to as fixed (individual) effects.

# Introduction Contd...

- $\alpha_i$  captures all unobservable time-invariant differences across individuals
- The FE doesn't require that  $\alpha_i$  &  $x_{it}$  to be uncorrelated
- As we'll see in a short while, the assumption of  $\alpha_i$  as fixed parameters has some great advantages but also some disadvantages.



# Introduction Contd...

## Efficiency of Parameters Estimators

- Panel data are larger than cross-sectional and time-series data with variation over two dimensions (individuals and time)  $\implies$  estimators are more accurate.
- Even with identical sample sizes, the use of panel data will often yield more efficient estimators than a series of independent cross-sections,

# Introduction Contd...

## Identification of Parameters

- Panel data reduces identification problems.
  - Identification in the presence of endogenous regressors or measurement error
  - Robustness to omitted variables
  - Identification of individual dynamics

# The Fixed Effects Model

- A linear regression model in which the intercept terms vary over the individual units  $i$ , i.e.

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}, \quad u_{it} \sim IID(0, \sigma_u^2), \& E(x'_{it}u_{it}) = 0. \quad (4)$$

- One can rewrite this model in the usual way by introducing a dummy variable for each unit  $i$ ,

$$y_{it} = \sum_{j=1}^n \alpha_j d_{ij} + x'_{it}\beta + u_{it}, \quad (5)$$

- Where  $d_{ij} = 1$  if  $i = j$  & 0 otherwise

# The Fixed Effects Model Contd...

- We thus have  $N$  dummy variables in the model. The parameters  $\alpha_1, \dots, \alpha_N$  &  $\beta$  can be estimated by OLS
- The implied estimator for  $\beta$  is referred to as the *Least Square Dummy Variable* (LSDV) estimator.
- It may however be numerically inconvenient to have so many dummy variables in a regression model. Why?
- It is possible to estimate  $\beta$  in a simpler way! This involves transformation of the data to eliminate  $\alpha_i$
- To do so, note that

$$\bar{y}_i = \alpha_i + \bar{x}_i' \beta + \bar{u}_i, \quad (6)$$

# The Fixed Effects Model Contd...

- Where  $\bar{y}_i = T^{-1} \sum_t y_{it}$  &  $\bar{x}_i$  &  $\bar{u}_i$  are defined in a similar way. Consequently, one can write

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)' \beta + (u_{it} - \bar{u}_i) \quad (7)$$

- You can see that  $\alpha_i$  is gone!
- The transformation that produces equation (7) is called the *within transformation*
- The OLS estimator for  $\beta$  obtained from this transformed model is often called the *within or fixed effects estimator* and is exactly identical to the LSDV estimator described above.

## The Fixed Effects Model Contd...

- The within estimator is given by,

$$\hat{\beta}_{FE} = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \quad (8)$$

- For consistency, it is required that

$$E\{(x_{it} - \bar{x}_i)u_{it}\} = 0 \quad (9)$$

- Sufficient for this is that  $x_{it}$  is uncorrelated with  $u_{it}$ , and that  $\bar{x}_i$  has no correlation with the error term!  $\implies$

$$E\{x_{it}u_{is}\} = 0 \quad \text{for all } s, t. \quad (10)$$

- $\implies x_{it}$  are strictly exogenous.

## The Fixed Effects Model Contd...

- The covariance matrix for  $\hat{\beta}_{FE}$  assuming that  $u_{it}$  is i.i.d. across  $i$  &  $t$  with variance  $\sigma_u^2$ , is given by

$$V\{\hat{\beta}_{FE}\} = \sigma_u^2 \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \quad (11)$$

- Unless  $T$  is large, using OLS estimate for the covariance matrix based upon the within regression in equation (11) will underestimate the true variance. Why?
  - In the transformed regression the error covariance matrix is singular (as the  $T$  transformed errors of each individual add up to zero) and the variance of  $u_{it} - \bar{u}_i$  is  $(T-1)/T \sigma_u^2$  rather than  $\sigma_u^2$

## The Fixed Effects Model Contd...

- A consistent estimator for  $\sigma_u^2$  is obtained from the sum of squared residuals from the within estimator, divided by  $N(T - 1)$ .
- Defining

$$\hat{u}_{it} = y_{it} - \hat{\alpha}_i - x'_{it}\hat{\beta}_{FE} = y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i)' \hat{\beta}_{FE}, \quad (12)$$

- One can estimate  $\sigma_u^2$  as

$$\hat{\sigma}_u^2 = \frac{1}{N(T - 1)} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2. \quad (13)$$

- Note: the FE model concentrates on differences "within" individuals, not differences with each others.



# The First Difference Estimator

- An alternative way to illuminate  $\alpha_i$  is to first-difference Eq(4) as

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + (u_{it} - u_{i,t-1})$$

$$\Delta y_{it} = \Delta x'_{it}\beta + \Delta u_{it} \quad (14)$$

- Applying OLS to this equation gives the First Differences (FD) estimator

$$\hat{\beta}_{FD} = \left( \sum_{i=1}^N \sum_{t=2}^T \Delta x_{it} \Delta x'_{it} \right)^{-1} \sum_{i=1}^N \sum_{t=2}^T \Delta x_{it} \Delta y_{it}. \quad (15)$$

## The First Difference Estimator Contd...

- Consistency requires:  $E\{\Delta x_{it}\Delta u_{it}\} = 0$  or

$$E\{(x_{it} - x_{it-1})(u_{it} - u_{it-1})\} = 0 \quad (16)$$

- This is a weaker condition than the strict exogeneity condition in Eq (10).
- For eg. it would allow correlation between  $x_{it}$  and  $u_{it-2}$
- Computation of standard errors for  $\hat{\beta}_{FD}$  requires taking in to account serial correlation in  $\Delta u_{it}$
- Since the conditions for consistency of the FD estimator are slightly weaker than those for the FE estimator, it is in general, somewhat less efficient.
- For  $T = 2$ , both estimators are identical.
- If the two estimators provide very different results  $\implies$  assumption (10) is problematic!

# The Random Effects Estimator

- In this model,  $\alpha_i$  is assumed to be random factors, independently and identically distributed over individuals. The model is specified as

$$y_{it} = \beta_0 + x'_{it}\beta + \alpha_i + u_{it}, \quad u_{it} \sim IID(0, \sigma_u^2); \quad \alpha_i \sim IID(0, \sigma_\alpha^2) \quad (17)$$

$$\varepsilon_{it} = \alpha_i + u_{it} \quad (18)$$

- Also referred to as a *one-way error components model*.
- All correlation of  $\varepsilon_{it}$  is attributed to  $\alpha_i$
- It is assumed that  $\alpha_i$  and  $u_{it}$  are mutually independent and independent of  $x_{js}$ ,  $\forall j \& s$
- $\implies$  estimating Eq(17) by OLS results in unbiased and consistent parameter estimates.

# The Random Effects Estimator Contd...

- However,  $\varepsilon_{it}$  exhibits a particular form of autocorrelation (unless in a special case where  $\sigma_\alpha^2 = 0$ )
- Consequently, the standard errors computed using OLS are incorrect and a more efficient estimator that exploits the structure of the error covariance matrix can be computed using GLS.
- Let error terms for individual  $i$  be stacked as  $\alpha_i l_T + u_i$ , where  $l_T = (1, 1, \dots, 1)'$  of dimension  $T$  and  $u_i = (u_{i1}, \dots, u_{iT})'$ ,
- The Cov matrix of this vector

$$V\{\alpha_i l_T + u_i\} = \Omega = \sigma_\alpha^2 l_T l_T' + \sigma_u^2 I_T, \quad (19)$$

## The Random Effects Estimator Contd...

- Where  $I_T$  is the T-dimensional identity matrix
- Premultiply the vectors  $y_i = (y_{i1}, \dots, y_{iT})'$ , etc., by  $\Omega^{-1}$ , which is given by

$$\Omega^{-1} = \sigma_u^{-2} \left[ I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2} l_T l_T' \right] \quad (20)$$

- or

$$\Omega^{-1} = \sigma_u^{-2} \left[ (I_T - 1/T * l_T l_T') + \psi 1/T * l_T l_T' \right], \quad (21)$$

- Where

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2} \quad (22)$$

## The Random Effects Estimator Contd...

- Note that  $I_T - (1/T)l_Tl_T'$  transforms the data in deviations from individual means and  $(1/T)l_Tl_T'$  takes individual means.
- The GLS estimator for  $\beta$  can be given by

$$\hat{\beta}_{GLS} = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' + \psi T \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right)^{-1} \\ \times \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) + \psi T \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) \right) \quad (23)$$

- Where  $\bar{x}$  refers to overall average of  $x_{it}$
- If  $T \implies \infty$ ,  $\psi \implies 0 \implies$ , the FE estimator arises (FE & RE become equivalent).
- If  $\psi = 1$ , the GLS estimator is just the OLS estimator (and  $\Omega$  is diagonal).

## The Random Effects Estimator Contd...

- From the general formula for the GLS, one can derive:

$$\hat{\beta}_{GLS} = W\hat{\beta}_B + (I_K - W)\hat{\beta}_{FE}, \quad (24)$$

- Where,

$$\hat{\beta}_B = \left( \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right)^{-1} \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}), \quad (25)$$

- Is called the *between estimator* for  $\beta$ .
- Is nothing but the OLS estimator in the model for individual means

$$\bar{y}_i = \beta_0 + \bar{x}_i' \beta + \alpha_i + \bar{u}_i, \quad i = 1, \dots, N. \quad (26)$$

# The Random Effects Estimator Contd...

- The matrix  $W$  is a weighting matrix and is proportional to the inverse of the covariance matrix of  $\hat{\beta}_B$
- The GLS estimator:
  - is a matrix-weighted average of the between estimator and the within estimator, where the weight depends upon the relative variances of the two estimators.
  - Under the current assumptions, is the optimal combination of the within estimator and the between estimator, and is therefore more efficient than either of these two estimators



# The Random Effects Estimator Contd...

- The OLS estimator (with  $\psi = 1$ ) is also a linear combination of these two estimators, but not the efficient one.
- The GLS estimator is more efficient than OLS and if the explanatory variables are independent of all  $u_{it}$  and all  $\alpha_i$ , it is unbiased.
- It will be consistent if in addition to (9) it also holds that  $E\{\bar{x}_i u_{it}\} = 0$ , and most importantly that

$$E\{\bar{x}_i \alpha_i\} = 0 \quad (27)$$

- These conditions are also required for the between estimator to be consistent.

## The Random Effects Estimator Contd...

- The covariance matrix of the RE estimator is given by

$$V\{\hat{\beta}_{RE}\} = \sigma_u^2 \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' + \psi T \sum_{i=1}^N (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right) \quad (28)$$

- since it makes use of the between variation in the data  $(\bar{x}_i - \bar{x})$  the RE estimator is more efficient than the FE estimator as long as  $\psi > 0$ .

## FE or RE?

- The choice between the two is not easy, and in many applications, particularly when  $T$  is small, the differences in the estimates for  $\beta$  can be substantial.
- The common approach is to use *Hausman's test* - a test for  $H_0 : E\{x_{it}, \alpha_i\} = 0$
- A significant difference in the two estimators implies that the null hypothesis is unlikely to hold.
- Assume that  $E\{u_{it}x_{is} = 0\}$  for all  $s, t$ , so that  $\hat{\beta}_{FE}$  is consistent for  $\beta$  whether  $x_{it}$  &  $\alpha_i$  are uncorrelated, whereas  $\hat{\beta}_{RE}$  is consistent and efficient only if  $x_{it}$  and  $\alpha_i$  are not correlated.
- Hausman's test evaluates the difference between the two estimators  $(\beta_{FE} - \beta_{RE})$ .
- It is easy to show that under the null,

$$V\{\hat{\beta}_{FE} - \hat{\beta}_{RE}\} = V\{\hat{\beta}_{FE}\} - V\{\hat{\beta}_{RE}\} \quad (29)$$

## FE or RE? Contd...

- Thus, the Hausman statistic is given by

$$\xi_H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' [\hat{V}\{\hat{\beta}_{FE}\} - \hat{V}\{\hat{\beta}_{RE}\}]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \quad (30)$$

- Where the  $\hat{V}_s$  denote estimates of the true covariance matrices
- Under the null hypothesis,  $(plim)(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = 0$ , the statistic  $\xi_H$  has an asymptotic Chi-squared distribution with K degrees of freedom, where K is the number of elements in  $\beta$

## Hausman - Taylor Estimator

- The FE model eliminates anything that is time invariant (E.g, Gender) from the model
- Imposing the assumption that  $E\{x_{it}, \alpha_i\} = 0$  to deal with the above problem might be unjustifiable
- The way out is to use an IV method which is considered to be in between the FE & RE approaches
- Re-write the FE model as:

$$\begin{aligned}\hat{\beta}_{FE} &= \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \\ &= \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)x_{it}' \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)y_{it}.\end{aligned}\quad (31)$$

## Hausman - Taylor Estimator Contd...

- This expression has the interpretation of an IV estimator for  $\beta$  in the model

$$y_{it} = \beta_0 + x'_{it}\beta + \alpha_i + u_{it} \quad (32)$$

- Each explanatory variable is instrumented by its value in deviation from the individual specific mean.
- I.e.,  $x_{it}$  is instrumented by  $x_{it} - \bar{x}_i$ .
- $E\{x_{it} - \bar{x}_i\}\alpha_i\} = 0$  by construction, and hence the IV estimator is consistent if it holds that  $E\{x_{it} - \bar{x}_i\}u_{it}\} = 0 \implies x_{it}$  is strictly exogenous.

## Hausman - Taylor Estimator Contd...

- To elaborate the general approach of the HT estimator, consider a linear model with four groups of explanatory variables:

$$y_{it} = \beta_0 + x'_{1,it}\beta_1 + x'_{2,it}\beta_2 + w'_{1i}\gamma_1 + w'_{2i}\gamma_2 + \alpha_i + u_{it}, \quad (33)$$

- Where the  $x$  vars are time varying and the  $w$  vars are time invariant
- The variables with index 1 are assumed to be uncorrelated with both  $\alpha_i$  &  $u_{it}$ , while the ones with index 2 are correlated with  $\alpha_i$  but not with  $u_{it}$

# Hausman - Taylor Estimator Contd...

- Hausman and Taylor show that equation (33) can be estimated by instrumental variables using the following variables as instruments:  $x_{1,it}$ ,  $w_{1i}$  &  $x_{2,it} - \bar{x}_{2i}$ ,  $\bar{x}_{1i}$ .
- Note:
  - The exogenous variables serve as their own instruments
  - $x_{2,it}$  is instrumented by its deviation from individual means (as in the FE approach)
  - $w_{2i}$  is instrumented by the individual average of  $x_{1,it}$
  - We don't need to use external instruments! (an attractive advantage of the HT model)
- Identification requires that the number of variables in  $x_{1,it}$  is at least as large as that in  $w_{2i}$ .



# Linear Panel Data Models - Summary

- The between estimator
  - Exploits the between dimensions of the data (differences between individuals)
  - Uses OLS in a regression of individual averages of  $y$  on individuals averages of  $x$  (and a constant)
  - Consistency as  $N \rightarrow \infty$  requires that  $E\{\bar{x}_i \alpha_i\} = 0$
- The FE (within) estimator
  - Exploits the within dimension of the data (differences within individuals)
  - Applies OLS in a regression in deviations from individual means
  - Consistency for  $\beta$  for  $T \rightarrow \infty$  or  $N \rightarrow \infty$  requires  $E\{(x_{it} - \bar{x}_i)u_{it}\} = 0$
  - Does not impose  $E\{(x_{it} - \bar{x}_i)\alpha_i\} = 0$

## Linear Panel Data Models - Summary Contd...

- The OLS estimator
  - Exploits both dimensions (within and between)
  - Not efficient
  - Consistency for  $T \rightarrow \infty$  or  $N \rightarrow \infty$  requires that  $E\{x_{it}(u_{it} + \alpha_i)\} = 0$
  - Doesn't impose strict exogeneity
- The RE (GLS) estimator
  - Combines the information from the between and within dimensions in an efficient way
  - Consistent under the combined conditions of the between and within estimators
  - Can be determined as a weighted average of the between and within estimator

# Linear Panel Data Models - Summary Contd...

- The FD estimator
  - OLS after first- differencing the equation of interest
  - Can serve as an alternative to the FE estimator
  - Consistency requires that  $E\{(x_{it} - x_{i,t-1})(u_{it} - u_{i,t-1})\} = 0$
  - If  $u_{it}$  is i.i.d., the FD estimator is less efficient than the FE estimator; if  $T = 2$ , they are identical