Econometrics

Week 6

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Fall 2021

Recommended Reading

For today

- Simple panel data methods
- second part of Chapter 13
- Advanced Panel Data Methods.
- Chapter 14

Next week

- Instrumental Variables Estimation and Two Stage Least Squares
- Chapter 15

Panel Data

- For a cross-section of individuals, schools, firms, cities, etc., we have several periods of data.
- Data are not independent, as in pooled cross-sections, the same individuals are observed in each time period.
- This means we might face similar problems as with time series data! e.g. autocorrelation
- This also means we can take advantage of panel structure of the data and use it to solve some kinds of omitted variable bias.
- To see this, let us write a model capturing the panel structure of the data

General Model for Panel Data

Unobserved Effects Model (Fixed Effects Model)

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \underbrace{a_i + u_{it}}_{\nu_{it}}$$

- ν_{it} is the **composite error**. It consists of
  - Time-invariant, individual specific, **unobserved effect** a_i
  - Time and individual specific **idiosyncratic error** u_{it}
- a_i is also referred to as **unobserved heterogeneity**, or individual heterogeneity, or **fixed effect**, because it is fixed over time.

Note that some variables are time variant and some time invariant (don't have the t-index)

Panel Data

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \underbrace{a_i + u_{it}}_{\nu_{it}}$$

- It is tempting to estimate this model by pooled OLS, but...
- It will be inefficient if errors are serially correlated
  - ...and they are correlated, because a_i is repeated every period
- It will be biased and inconsistent if u_{it} and x_{it} are correlated $\Rightarrow$ endogeneity bias that can be met also in cross-sectional models
- It will be biased and inconsistent if a_i and x_{it} are correlated: $Cov(a_i, x_{it}) \neq 0 \Rightarrow$ **heterogeneity bias**.

Panel Data

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \underbrace{a_i + u_{it}}_{\nu_{it}}$$

- There are several basic panel data methods developed to deal with specific panel data issues.
  - **Random effects transformation** used when $E(a_i|\mathbf{X}_i) = 0$ and $E(u_{it}|\mathbf{X}_i, a_i) = 0$. This approach makes the most efficient use of the panel data structure.
  - **Fixed effects transformation** used when $E(a_i|\mathbf{X}_i) \neq 0$, but $E(u_{it}|\mathbf{X}_i, a_i) = 0$. This approach removes a_i from the model and thus deals with heterogeneity bias.
  - **First differencing** used when $E(a_i|\mathbf{X}_i) \neq 0$, but $E(u_{it}|\mathbf{X}_i, a_i) = 0$. This approach also removes a_i from the model and thus deals with heterogeneity bias.

First-differenced estimator

First-differenced estimator (FD)

Let us start with two-period panel data.

$$\begin{aligned} y_{i2} &= (\beta_0 + \delta_0) + \beta_1 x_{i2} + \beta_2 z_i + a_i + u_{i2}, \quad (t = 2) \\ y_{i1} &= \beta_0 + \beta_1 x_{i1} + \beta_2 z_i + a_i + u_{i1}, \qquad (t = 1) \end{aligned}$$

Subtracting second equation from the first one gives:

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- Here, a_i is *"differenced away"*.
- Note that as a side effect z_i is also "differenced away"
- Can we estimate this equation by OLS and get a reliable estimate of β_1?

First-differenced estimator

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \underbrace{a_i + u_{it}}_{\nu_{it}}$$

$$\downarrow$$

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- Differencing is a powerful way to deal with time constant unobserved effects
- However, first-differencing can greatly reduce variation in the explanatory variables, and
- First-differencing removes observed time-constant variables from the regression.
- OLS estimates of parameters in the first-differenced equation are unbiased as long as the following assumptions are satisfied:

Assumptions for Pooled OLS Using First Differences

Assumption FD1

For each observation i, the model is
$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, \ldots, T,$
where parameters β_j are to be estimated and a_i is the unobserved fixed effect.

Assumption FD2

Each period we observe the same random sample.

Assumption FD3

Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables.

Assumptions for Pooled OLS Using First Differences

Assumption FD4

Let $\mathbf{X}_i$ denote x_{itj}, $t = 1, \ldots, T$, $j = 1, \ldots, k$. For each t, the expected value of the idiosyncratic error given the explanatory variables in *all* time periods and the unobserved effect is zero: $E(u_{it}|\mathbf{X}_i, a_i) = 0$.

An important implication of FD4 is that $E(\Delta u_{it}|\mathbf{X}_i) = 0$, $t = 2, \ldots, T$. Once we control for a_i, there is no correlation between the x_{isj} and the remaining error u_{it} for all s and t. x_{itj} is strictly exogenous conditional on the unobserved effect.

- Under assumptions FD1 - FD4, the first-difference estimator is unbiased.

Assumptions for Pooled OLS Using First Differences

Assumption FD5

The variance of the differenced error, conditional on all explanatory variables, is constant: $Var(\Delta u_{it}|\mathbf{X}_i) = \sigma^2$, for all $t = 2, \ldots, T$.

Assumption FD6

For all $t \neq s$, the differences in the idiosyncratic errors are uncorrelated (conditional on all explanatory variables): $Cov(\Delta u_{it}, \Delta u_{is}|\mathbf{X}_i) = 0$, $t \neq s$.

- Under assumptions FD1 - FD6, the first-difference estimator is BLUE.

Assumptions for Pooled OLS Using First Differences

Assumption FD7

Conditional on $\mathbf{X}_i$, the Δu_{it} are independent and identically distributed normal random variables.

- This last assumptions assures that FD estimator is normally distributed, t and F statistics from the pooled OLS on the differenced data have exact t and F distributions.

Differencing with More than Two Periods

- We can extend FD to more than two periods.
- We simply difference adjacent periods.

A general fixed effects model for N individuals and t=1,2,3

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$$

- The total number of observations is 3N.
- The key assumption is that idiosyncratic errors are uncorrelated with explanatory variables: $Cov(x_{itj}, u_{is}) = 0$ for all t, s and $j \Rightarrow$ **strict exogeneity**.
- How to estimate? Simply difference equation for $t = 1$ from $t = 2$ and $t = 2$ from $t = 3$.
- It will result in 2 equations which can be estimated by pooled OLS consistently under the CLM assumptions.
- We can simply further extend to T periods.
- Correlation and heteroskedasticity are treated in the same way as in time series data.

Fixed Effects Transformation

Let us go back to the unobserved effects regression model

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}, \qquad t = 1 \dots T.$$

- When unobserved heterogeneity, represented by a_i is correlated with explanatory variable x_{it}, OLS estimates of this model parameters are **biased** and **inconsistent**.
- First differencing is one of the ways how to eliminate the unobserved effect a_i and obtain unbiased, consistent estimates of model parameters.
- An alternative, which is more efficient when u_{it} is well-behaved, is called **the fixed effects transformation**.

Fixed Effects Transformation

- Take the unobserved effects model

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}, \qquad t = 1 \ldots T, \quad i = 1 \ldots n$$

- For each i, average the equation over time:

$$\overline{y}_i = \beta_0 + \beta_1 \overline{x}_i + a_i + \overline{u}_i,$$

where $\overline{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$, and similarly for $\overline{x}_i$ and $\overline{u}_i$.

- Subtracting the averages from the original equation, we get **time-demeaned model**:

$$\begin{aligned}
\underbrace{y_{it} - \overline{y}_i}_{\ddot{y}_{it}} \quad &= \quad \beta_0 - \beta_0 + \beta_1 (\underbrace{x_{it} - \overline{x}_i}_{\ddot{x}_{it}}) + a_i - a_i + \underbrace{u_{it} - \overline{u}_i}_{\ddot{u}_{it}}, \\
\ddot{y}_{it} \quad &= \quad \beta_1 \ddot{x}_{it} + \ddot{u}_{it} \qquad t = 1 \ldots T, \quad i = 1 \ldots n
\end{aligned}$$

Fixed Effects Transformation

- This fixed effects transformation is also called the **within transformation**.
- Unobserved effect a_i disappeared $\Rightarrow$ omitted variable bias is no longer a problem $\Rightarrow$ we can use pooled OLS.
- Pooled OLS estimator using time-demeaned variables is called **the fixed effects (FE) estimator**, or **the within estimator**.
- The name *"within"* comes from the fact that we use time variation **within** each cross-sectional observation.

- We also know a between estimator, which is obtained using the OLS estimation of $\overline{y}_i = \beta_0 + \beta_1 \overline{x}_i + a_i + \overline{u}_i$.
- We use time-averages and then run a cross-sectional regression.
- Between estimator is biased when a_i is correlated with x_i.

Fixed Effects Transformation

- A general time-demeaned equation for each cross-sectional unit i is:

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} + \ldots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it},$$

  for $t = 1, 2, \ldots, T$, and we estimate it by pooled OLS.
- Note that the intercept is eliminated by the fixed effects transformation.
- Let us discuss the necessary assumptions and properties of the fixed effects estimator, $\hat{\beta}_{FE}$.

Assumptions for Fixed Effects Estimator

Assumption FE1

For each i, the model is
$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, \ldots, T, , \quad i = 1 \ldots n,$
where parameters β_j are to be estimated and a_i is the unobserved fixed effect (unebserved heterogeneity).

Assumption FE2

We have a random sample in the cross-sectional dimension.

Assumption FE3

Each explanatory variable changes over time (for at least some i), and there are no perfect linear relationships among the explanatory variables.

Assumptions for Fixed Effects Estimator

Assumption FE4

For each t, the expected value of the idiosyncratic error given the explanatory variables in *all* time periods and the unobserved effect is zero: $E(u_{it}|\mathbf{X}_i, a_i) = 0$.

- Under Assumptions FE1-FE4 (which are **identical** as for the first-differencing estimator), $\hat{\beta}_{FE}$ is unbiased. The key assumption is strict exogeneity (FE4).
- Under Assumptions FE1-FE4, $plim(\hat{\beta}_{FE}) = \beta$ as $N \to \infty$ ($\hat{\beta}_{FE}$ is consistent).

Assumptions for Fixed Effects Estimator

Assumption FE5

$Var(u_{it}|\mathbf{X}_i, a_i) = Var(u_{it}) = \sigma_u^2$, for all $t = 1, \ldots, T$.

Assumption FE6

For all $t \neq s$, the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and a_i):
$Cov(u_{it}, u_{is}|\mathbf{X}_i, a_i) = 0$.

- Under the Assumptions FE1-FE6, the fixed effects estimator is BLUE.

Assumptions for Fixed Effects Estimator

Assumption FE7: Normality

Conditional on $\mathbf{X}_i$ and a_i, the u_{it} are independent and identically distributed as Normal$(0, \sigma^2)$.

- Assumption FE7 assures us that FE estimator is normally distributed, its t and F statistics have exact t and F distributions respectively.
- Without FE7, we can rely on asymptotic approximations (although, without further assumptions, they require large N and small T).

Least Squares Dummy Variable Estimator

- An estimator numerically identical to the FE estimator can be obtained by adding individual dummies for each cross-sectional observation
(to estimate unobserved effect for each i individually).
- This **dummy variable regression** necessarily has many explanatory variables ⇒ dummy variables are often not practical but sometimes we are interested in the estimation of individual fixed effects.
- Careful! While LSDV regression produces consistent estimates of β's, $\hat{a}_i$'s are inconsistent!
    - With fixed T, as $N \to \infty$.
- Using dummy variable regression we can see better why variables that are constant over time cannot be used in FE regression.

Fixed Effects (FE) vs. First Differencing (FD)

- FD involves **differencing** the data, FE involves **time-demeaning**. Which one to use?
- FD and FE estimates are *identical* when $T = 2$.
- For $T > 2$, the methods are different.
- If u_{it} is serially uncorrelated, FE is more efficient than FD.
- If u_{it} follows a Random Walk, then Δu_{it} is serially uncorrelated and FD is better $\Rightarrow$ test whether Δu_{it} are serially correlated first.
- But in most of the data serial correlation is not that strong as in Random Walk.
- Thus it is suggested to obtain both estimates. If the results are not sensitive, then it is fine. But if they vary, we have to find out why!
- Careful when T is large and N is small!

Random Effects Models

- In FE or FD estimation, we would like to eliminate a_i because we expect it is correlated with x_{itj}.
- Now, suppose that a_i is *uncorrelated* with each explanatory variable, indexed by j, at all periods, indexed by t.
- Are FE and FD efficient?
  No, because we eliminate the information in a_i.
- Solution is to use **Random Effects Model.**

Random Effects Model (RE)

$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$,
where $Cov(x_{itj}, a_i) = 0$
for all $t = 1, 2, \ldots, T$ and $j = 1, 2, \ldots, k$.

Random Effects Models

- Note that if a_i is uncorrelated with explanatory variables, single cross-section OLS is **consistent**.
- Thus we may not need panel data at all.
- If a_i is uncorrelated with explanatory variables, Pooled OLS is also **consistent**.
- But, in this case, we throw away useful information.
  - We know that observations within cross-sectional units share common unobserved characteristics.
  - Random errors are serially correlated!
  - Pooled OLS is not efficient.

Random Effects Models

- Let us consider the following regression equation:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \nu_{it},$$

  where $\nu_{it} = a_i + u_{it}$ is a **composite error term**.
- The ν_{it} are serially correlated across time, as a_i is present in each time period.

$$Corr(\nu_{it}, \nu_{is}) = \sigma_a^2/(\sigma_a^2 + \sigma_u^2),$$

for all $t \neq s$, where $\sigma_a^2 = Var(a_i)$ and $\sigma_u^2 = Var(u_{it})$

- Because of this positive serial correlation, pooled OLS estimator is inefficient (and gives wrong standard errors).

Random Effects Models

- Solution to this problem is a GLS transformation that eliminates serial correlation in the errors
(like in the case of serial correlation in time series models).
- Transformation subtracts a fraction of time average, where the fraction depends on σ_u^2, σ_a^2, and the number of time periods T:

$$\begin{aligned} y_{it} - \lambda \overline{y}_i &= \beta_0(1-\lambda) + \beta_1(x_{it1} - \lambda \overline{x}_{i1}) + \ldots \\ &\quad + \beta_k(x_{itk} - \lambda \overline{x}_{ik}) + (\nu_{it} - \lambda \overline{\nu}_i) \end{aligned}$$

where $\lambda = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)]^{1/2}$ and $\overline{y}_i$ is time average.

- This equation contains **quasi-demeaned data**.
- Errors are now uncorrelated, and the GLS estimator is simply the pooled OLS of this transformation.

Random Effects Models

- Advantage of RE is that it allows for explanatory variables which are constant over time (as opposed to FE).
- In practice, λ is never known, as it is composed of theoretical variances.
- We need to estimate it, usually by Pooled OLS: $\hat{\lambda} = 1 - [\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + T\hat{\sigma}_a^2)]^{1/2}$, where $\hat{\sigma}_u^2$ and $\hat{\sigma}_a^2$ are consistent estimators of σ_u^2 and σ_a^2 respectively under Pooled OLS.
- Thus, random effects model is estimated by **feasible GLS** – FGLS, where λ is replaced by $\hat{\lambda}$.

Note

- For $\lambda = 0$, we have Pooled OLS (a_i is unimportant as it has small variance relative to u_{it}).
- For $\lambda = 1$, we have FE (σ_a^2 is large relatively to σ_u^2).

Assumptions for Random Effects

- Assumptions FE1, FE2 and FE6 are the same for the RE model.

RE3

There are no perfect linear relationships among the explanatory variables.

$\Rightarrow$ allow explanatory variables to be constant in time for all i.

RE4

In addition to FE4, the expected value of a_i given all explanatory variables is constant: $E(a_i|\mathbf{X}_i) = 0$.

$\Rightarrow$ rule out correlation between unobserved effect and explanatory variables.

Assumptions for Random Effects

RE5

In addition to FE5, the variance of a_i given all explanatory variables is constant: $Var(a_i|\mathbf{X}_i) = \sigma_a^2$.

- Under Assumptions FE1, FE2, RE3 and RE4, the random effects estimator $\hat{\beta}_{RE}$ is consistent as N gets large for fixed T.
- RE estimator is not unbiased unless we know λ.
- Under the FE1, FE2, RE3, RE4, RE5 and FE6, the RE estimator is also approximately asymptotically normally distributed with large N and usual standard errors, t statistics and F statistics are valid.

Fixed Effects vs. Random Effects

- We decide whether to use RE or FE based on a_i.
- If unobserved effect is something we want to estimate, use FE.
- If unobserved effect is supposed to be random, use RE.
- But, to treat a_i as random, we have to make sure that it is not correlated with explanatory variables.
- If unobserved effect a_i is correlated with explanatory variables, FE is consistent, while **RE is inconsistent**.
- Otherwise, RE is more efficient than FE.

Fixed Effects vs. Random Effects

- We can test statistically whether to use FE or RE:

Hausman test

- $H_0 : Cov(a_i, x_{it}) = 0$.
- Under the null, **both FE and RE are consistent**, but **RE is asymptotically more efficient**.
- Under the alternative, FE is still consistent (RE is not).

- We can test and correct for serial correlation and heteroskedasticity in the errors.
- We can estimate standard errors robust to both.

Thank you

Thank you for your attention!

... and do not forget to read Chapter 15 for the next week!

Remember about the Home Assignment - Due November 12

Midterm exam on November 18 at 9am - ONLINE!!!