Panel Data Methods

BS1802 Statistics and Econometrics

Jiahua Wu

382 Business School
j.wu@imperial.ac.uk

Imperial College Business School

Imperial means Intelligent Business

Roadmap

- Regression analysis with cross-sectional data
  - The multiple regression analysis
    - Basics: estimation, inference, analysis with dummy variables
    - More technically involved: asymptotics, heteroskedasticity, specification and data issues
- Advanced topics
  - Limited dependent variable models
  - Panel data analysis
  - Regression analysis with time series data

Outline (Wooldridge, Ch. 13.3, 13.5, 14.1)

- Two period panel data
- First-differenced estimation
- Fixed effects estimation

Outline

- Two period panel data
- First-differenced estimation
- Fixed effects estimation

What is Panel Data?

- A set of panel data
  - has both a cross-sectional and a time series dimension
  - is collected by following the same individuals over a number of time periods
- E.g., a panel data set for *wage*, *educ*, *exper*, ...
  1. Randomly select a sample of people from the population and collect data for 2016
  2. The same people are re-interviewed to collect data for 2017, 2018, ...
- It is possible to use a panel just like cross sections, but can do more than that
- Panel data allows us to address issues related to unobserved factors, which are difficult to handle with cross sectional data

Two Period Panel Data

- Example 9.4. City Crime Rates
  - Data: crime rates (*crmrte*) and unemployment rates (*unem*) from a sample of 46 cities in 1982 ($t = 1$) and 1987 ($t = 2$).
  - Question: Did *unem* influence *crmrte*?
  - Regressing *crmrte* on *unem* using the sample from 1987, we have

$$\widehat{crmrte}_{87} = \underset{(20.76)}{128.38} - \underset{(3.42)}{4.16}\, unem_{87},$$

  $n = 46, R^2 = .033$

  - The result is likely biased because many relevant factors (e.g., city, police, ...) are not controlled for

Two Period Panel Data

- An alternative way to look at the data
  - If the omitted variables are fixed over time, then we can decompose the error into two parts: factors that vary over time and those do not
- Consider the previous example in the panel setting

$$crmrte_{it} = \beta_0 + \delta_0 d2_t + \beta_1 unem_{it} + a_i + u_{it}, \quad t = 1, 2$$

where

  - i is the city
  - t is the time period
  - $d2_t$ is the dummy variable indicating the second time period
  - A time-constant component is added to the error $v_{it} = a_i + u_{it}$

Fixed Effects Model

- In general, the fixed-effects model can be written as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, 2$$

  where

  - a_i is the fixed effect (invariant to t) that represents factors specific to individual i (allowed to be correlated with $\mathbf{x}_{it}$)
  - u_{it} is called the idiosyncratic error that represents unobserved factors varying both overtime and across sections (typically assumed to be uncorrelated with $\mathbf{x}_{it}$)

Outline

- Two period panel data
- First-differenced estimation
- Fixed effects estimation

First-Differenced Estimation

- Write the model separately

$$y_{i1} = \beta_0 + \delta_0 \cdot 0 + \beta_1 x_{i11} + \cdots + \beta_k x_{i1k} + a_i + u_{i1}, \quad (t = 1)$$
$$y_{i2} = \beta_0 + \delta_0 \cdot 1 + \beta_1 x_{i21} + \cdots + \beta_k x_{i2k} + a_i + u_{i2}, \quad (t = 2)$$

- Subtracting the first equation from the second one gives

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + \cdots + \beta_k \Delta x_{ik} + \Delta u_i,$$

  (first-differenced equation) which is a cross-section model and is free of a_i

- When u_{it} is uncorrelated with regressors in both periods
  - There is no correlation between Δx_i's and Δu_i
  - OLS will be unbiased

Panel Data Estimation in R

- The command to perform panel data estimation in R is `plm(formula, data, effect, model, index, …)`
  - `effect`
    - fixed effects for cross-sectional units (“`individual`”)
    - time effects (“`time`”)
    - both (“`twoways`”)
  - `model`
    - first-differences (“`fd`”)
    - fixed effects (“`within`”)
    - random effects (“`random`”)

First-Differenced Estimation: An Example

- Example 9.7. City Crime Rates.
  - First-differenced estimation

$$\Delta\widehat{crmrte} = \underset{(4.70)}{15.40} + \underset{(.88)}{2.22}\Delta unem,$$

$n = 46, R^2 = .127$

- There is a positive and significant relationship between $unem_{it}$ and $crmrte_{it}$
- One percentage point rise in unemployment rate increases 2.22 crimes per 1,000 people
- The crimes per 1,000 people increased by 15.4 in 1987, in comparison to 1982

First-Differenced Estimation: A Shortcoming

- Consider the log wage model with two-period panel

$$\log(wage_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + a_i + u_{it}, \quad t = 1, 2,$$

where a_i represents unobserved factors, say $ability_i$

  - The first-differenced equation is

$$\Delta \log(wage_i) = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i$$

  - However, for most adult workers, $\Delta educ_i$ is zero. The overall variation in $\Delta educ_i$ is small, and thus OLS estimator will have a large standard error
  - Using the first-differenced estimation is a good idea for "returns to eduction". But, frequently, it does not work well because of the lack of variation in $\Delta educ_i$

Panel Data with More than Two Periods

- For the panel data with T periods
  1. Subtract period 1 from period 2

     $\vdots$

  2. Subtract period $(T-1)$ from period T
  3. We have $(T-1)$ observations per individual
  4. Estimate by OLS, assuming the Δu_{it} are uncorrelated over time
- The key assumption about the idiosyncratic error u_{it} is

$$Cov(x_{itj}, u_{is}) = 0$$

- When using more than two time periods, we must assume that Δu_{it} is uncorrelated over time for the usual standard errors and test statistics to be valid
  - To deal with serial correlation in Δu_{it}, GLS method may be used

Outline

- Two period panel data
- First-differenced estimation
- Fixed effects estimation

Fixed Effects Estimation

- When there is an unobserved fixed effect, an alternative to first differences is fixed effects estimation
- Consider a model with a single explanatory variable

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

  - The average over time for individual i is

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

  - The average of a_i will be a_i. So if we subtract the average from y_{it}, we have

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i)$$

  - Each individual has been "de-meaned" for all variables, which eliminates the fixed effects.

First Difference vs Fixed Effects?

- When $T = 2$, first difference and fixed effects estimators will be exactly the same
- For $T > 2$
  - Both are unbiased (with T fixed as $N \to \infty$)
  - The relative efficiency of the estimators is determined by the serial correlation in u_{it}
    - When u_{it} are serial uncorrelated, fixed effects is typically more efficient
    - Serial correlation tests in `R`: `pwartest()`

Fixed Effects Estimation: An Example

- Example 14.1. Effect of Job Training on Firm Scrap Rates

$$\log(scrap_{it}) = \beta_0 + \delta_0 d88 + \delta_1 d89 + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + a_i + v_{it}$$

- Data description
  - 54 firms reported scrap rates in each of the three years, 1987, 1988, 1989
  - No firms received grants prior to 1988
  - In 1988, 19 firms received grants; in 1989, 10 different firms received grants
  - A lagged value of the grant indicator ($grant_{i,t-1}$) is included to allow for the possibility that the additional job training in 1988 made workers more productive in 1989

Fixed Effects Estimation: An Example

Dependent Variable: log(*scrap*)	
Independent Variables	**Coefficient (Standard Error)**
d88	−.080 (.109)
d89	−.247 (.133)
grant	−.252 (.151)
$grant_{-1}$	−.422 (.210)
Observations Degrees of freedom *R*-squared	162 104 .201