Advanced Methods Summer 21

Magdalene Silberberger

June 1, 2021

Panel Data

Outline

Feedback?

Introduction

Panel Data

- Key feature: Repeated observing of a unit over time
- Panel data allows us a researcher to study cross-section effects –i.e., along N, variation across the firms- time series effects –i.e., along T, variation across time
  1. Pooled OLS (POLS)
  2. Fixed Effects
  3. Random Effects

Panel Data: Advantages (Baltagi, 2014)

- allows you to control for *individual heterogeneity*
- more informative data, more variability, less collinearity among the dependent variables, more degrees of freedom and more efficiency in estimation
- identification and measurement of effects that are simply not detectable in pure cross-section or time-series data (e.g. more complicated behavioral models)
- reduction of biases resulting from aggregation over individual units

Panel Data: Limitations

- design and data collection problems:
  - coverage
  - non-response
  - frequency of interviewing (data collection in general)
- distortions of measurement errors
- selectivity problems:
  - self-selection
  - non-response
  - attrition
- short **T**
- cross-sectional dependence

Panel Data: Notation

$$y_{it} = \alpha + \beta_1 x_{1it} + ... + \beta_k x_{kit} + u_{it} \tag{1}$$

Panel Data: Notation

$$y_1 = \begin{bmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1t} \\ \dots \\ y_{1T} \end{bmatrix} ; \dots ; y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{it} \\ \dots \\ y_{iT} \end{bmatrix}$$

Panel Data: Notation

$$X_1 = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ x_{1t} & x_{2t} & \dots & x_{kt} \\ \dots & \dots & \dots & \dots \\ x_{1T} & x_{2T} & \dots & x_{kT} \end{bmatrix} \quad X_i = \begin{bmatrix} w_{11} & w_{21} & \dots & w_{k1} \\ w_{12} & w_{22} & \dots & w_{k2} \\ \dots & \dots & \dots & \dots \\ w_{1t} & w_{2t} & \dots & w_{kt} \\ \dots & \dots & \dots & \dots \\ w_{1T} & w_{2T} & \dots & w_{kT} \end{bmatrix}$$

Panel Data

A standard panel data set stacks the $y_i's$ and the $x_i's$:

$$y = X\beta + c + \epsilon \tag{2}$$

- **balanced:** has every observation from 1 to N observable in every period 1 to T
- **unbalanced:** has missing data

Pooled Data

Pooled Data

- you can simply run OLS with a sample of NT observations, ignoring the panel data structure
- you make no assumptions about individual differences

Problems:

- assumption: lack of correlation between errors corresponding to the same units
- if we relax the assumption: $cov(u_{i,t}, u_{i,s}) \neq 0$, we have *autocorrelation* and *heteroskedasticity*
- while our OLS estimator will still be consistent, the standard errors will be incorrect
- we could you *clustered/robust standard errors*, clustering on each individual unit

Fixed Effects

Fixed Effects

We can also relax the assumption that all individuals have the coefficients:

$$y_{it} = \alpha_i + \beta_1 x_{1it} + ... + \beta_k x_{kit} + u_{it} \tag{3}$$

- if our outcome variable depends on unobserved factors and the unobserved variables are correlated with the treatment, then our treatment variable is endogenous
- if these omitted variables that do not change over time we can use *fixed effects* to isolate a causal effect of x on y
- the easiest (but inefficient) way would be to include dummy variables (least square dummy variable estimator)

Fixed Effects

Alternatively, we can use the *Fixed Effects Estimator* that allows for individual intercepts (α_i) to control for *time invariant characteristics* and therefore capture individual heterogeneity

- We start with a simple fixed effects specification for unit i and subtract the average observation across time (entity demeaning)

Random Effects

Random Effects

- assumption: individual-specific effects are uncorrelated with the independent variables
- advantage: independent variables can be time invariant
- limitation: assumption often not met

Practice

Practice I

Use the *GUNS* dataset from the *AER package*

- load the dataset and get yourself an overview over the data
- estimate the model to empirically investigate the debate if and to what extent the right to carry a gun affects crime
- why does it make sense to add FE? (add FE if you haven't already and re-estimate the model)
- why would it make sense to add time FE? (add time FE and re-estimte the model)
- could there be a bias due to omitted socioeconomic characteristics? (add relevant ones)

Practice (II)

- choose one of the Wooldridge datasets (make sure it is a panel)
- think of a meaningful hypothesis that you can address with the data
- carefully think of control variables to avoid omitted variable bias
- do the regression(s) and interpret the results