Panel Data Methods

Econometrics II

Raquel Carrasco & Ricardo Mora

Department of Economics
Universidad Carlos III de Madrid
Máster Universitario en Desarrollo y Crecimiento Económico

Outline

A True Panel Data?

- up to now, we have seen two types of datasets:
  - cross sections: each observation represents an individual, firm, etc
  - time series: each observation represents a separate period
- we may also have cross sections for different time periods

Two Examples of Cross Sections with Time Effects

CIS surveys

- each month independently, the Centro de Investigaciones Sociológicas samples the Spanish population
- the surveys ask for political and sociological actitude

CPS

- each month independently, the US bureau of Labor Statistics quarterly samples the US population
- the Current Population Survey ask for economic activity and related issues

- data with this structure are known as pooled cross-sections
- you can take into account time effects

An Example of a True Panel

ECHP 1994–2001

- each year, the European Community Household Panel surveys the same individuals on a wide range of topics: income, health, education, housing, employment, etc
- in the first wave, i.e. in 1994, approximately 130,000 adults aged 16 years and over

- in panel data, we have observations for each individual for at least two consecutive periods
- also known as "longitudinal data"

Pooled Cross Sections

- we may want to pool cross sections just to get bigger samples
- we need to make assumptions about the value of the parameters in each period

we may assume that parameters remain constant

$$wages_{it} = \beta_0 + \beta_1 educ_{it} + u_{it}$$

- alternatively, we may want to investigate the effect of time

the simplest model assumes that only the intercept changes

$$wages_{it} = \beta_{0t} + \beta_1 educ_{it} + u_{it}$$

- this effectively controls for annual inflation

Time Variations in the Returns to Education (1/2)

we can investigate whether relations change in time

$$wages_{it} = \beta_0 + \beta_{1t} educ_{it} + u_{it}$$

- Step 1: generate a new variable which interacts x_{it} with year dummies
- Step 2a: run OLS of the dependent variable on all interactions plus a constant
  - each slope measures the returns each year
- Step 2b: run OLS on a constant and all interactions except one, say for the first year
  - each slope for the interactions measures how each year's returns differ from the first year

Time Variations in the Returns to Education (2/2)

What happens if we allow for parameter time variation in all years?

$$wages_{it} = \beta_{0t} + \beta_{1t} educ_{it} + u_{it}$$

- all beta coefficients estimated with a sample of N observations
- the standard deviation estimated with a sample of NT observations
- for the slopes, it is exactly like estimating all periods separately
- for inference, it is different

The Chow Test for Structural Change

$wages_{it} = \beta_{0t} + \beta_{1t} educ_{it} + u_{it}$

- suppose that there are two periods $t = 1, 2$
- $H_0 : \beta_{01} = \beta_{02}, \beta_{11} = \beta_{12}$
- compute an F test
- estimating the pooled regression is useful when we want the test to be robust to heteroskedasticity

An Example of Policy Analysis

effect on housing prices of building a garbage incinerator

- first suppose that there is only one period $t = 1981$

$$prices_{i,1981} = \beta_0 + \beta_1 near_{i,1981} + u_{i,1981}$$

- the hypothesis is that prices of houses located near the incinerator fall when the incinerator is announced/built

$$H_0 : \beta_1 = 0 \text{ vs } \beta_1 < 0$$

- $\hat{\beta}_1$ will be inconsistent if the incinerator is built in an area with lower housing prices: $cov(near_{i,1981}, u_{i,1981}) \neq 0$

Diff-in-diffs

suppose we also have data before announcement, say 1978

- $t = 1978, 1981$

$$prices_{it} = \beta_0 + \beta_1 near_{it} + \beta_3 D_{it}^{1981} + \beta_4 near_{it} D_{it}^{1981} + u_{i,t}$$

- the hypothesis is that prices of houses located near the incinerator fall when the incinerator is announced/built

$$H_0 : \beta_4 = 0 \text{ vs } \beta_4 < 0$$

- $\hat{\beta}_4$ is called the diff-in-diffs estimator
- it may consistently estimate the policy effect even if $cov(near_{i,1981}, u_{i,1981}) \neq 0$

When is the diff-in-diffs Estimator Consistent? (1/2)

$$E[prices|near, 1981] = \beta_0 + \beta_1 + \beta_3 + \beta_4$$

$$E[prices|near, 1978] = \beta_0 + \beta_1$$

First "diff": $E[prices_{1981} - prices_{1978}|near] = \beta_3 + \beta_4$

$$E[prices|far, 1981] = \beta_0 + \beta_3$$

$$E[prices|far, 1978] = \beta_0$$

Second "diff": $E[prices_{1981} - prices_{1978}|far] = \beta_3$

When is the diff-in-diffs Estimator Consistent? (2/2)

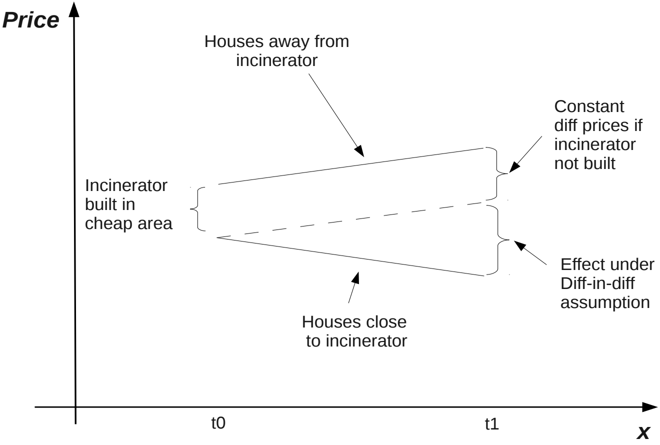
$$E[\Delta prices_{1981} \,|near\,] = \beta_3 + \beta_4$$

$$E[\Delta prices_{1981} \,|far\,] = \beta_3$$

"diff-in-diff": $E[\Delta prices_{near} - \Delta prices_{far}] = \beta_4$

- β_4 reflects the policy effect if the incinerator is built in an area with no "different inflation"

A Graphical Interpretation of Diff-in-Diff



Two-period Panel Data

In general, panel data can be used to address some kinds of omitted variable bias

$y_{it} = \beta_0 + \beta \mathbf{x}_{it} + a_i + u_{it}$

- a_i is a time-invariant, individual specific unobserved effect on the level of y
- if $cov(a_i, \mathbf{x}_{it}) \neq \mathbf{0}$ then $\hat{\beta}_0$ and $\hat{\beta}$ will be inconsistent

First Differences rids of the unobserved time-invariant components

$\Delta y_{it} = \beta \Delta \mathbf{x}_{it} + \Delta u_{it}$

- if $cov(\Delta u_{it}, \Delta \mathbf{x}_{it}) = \mathbf{0}$ then $\hat{\beta}_{FD}$ will be consistent

Two Examples of a First-Difference Estimator

$wages_{it} = \beta_0 + \beta_1 educ_{it} + \beta_2 IQ_i + u_{it}$

- taking first differences:$\Delta wages_{i2} = \beta_1 \Delta educ_{it} + \Delta u_{i2}$
- the sample is reduced: only individuals for whom there is a change in *educ* are used
- as long as $cov(\Delta educ, \Delta u) = 0$, $\hat{\beta}_1$ will be consistent
- $\hat{\beta}_1$ is called the first difference, *FD*, estimator

Twins data: $wages_{ij} = \beta_0 + \beta_1 educ_{ij} + \eta_i + u_{ij}$

- $j = 1$: older sibling
- $j = 2$: younger sibling

- to correctly compute the change, in STATA you can use *tsset*

Differencing with Many Periods

- we can extend *FD* to panel data with more than two periods
- simply difference adjacent periods

Three periods: $T = 3$

- take first differences twice

- estimate by OLS, assuming the Δu_{it} are uncorrelated over time

Fixed Effects

An alternative to first differences is substracting the individual specific mean

$y_{it} = \beta_0 + \beta \mathbf{x}_{it} + a_i + u_{it}$

- the model in averages: $\overline{y}_i = \beta_0 + \beta \overline{\mathbf{x}}_i + a_i + \overline{u}_i$

we get rid of a_i by substracting the mean

- $y_{it} - \overline{y}_i = \beta \left(\mathbf{x}_{it} - \overline{\mathbf{x}}_i \right) + \left(u_{it} - \overline{u}_i \right)$
- if $cov(\mathbf{x}_{it}, u_{is}) = \mathbf{0}$ then $\hat{\beta}_{FE}$ will be consistent

- the fixed effects estimator (FE) can be obtained by adding individual dummies to the regression

Properties of the *FE* Estimator

- under A1, A2 in the cross section, A4, and strict exogeneity on the controls, $plim(\hat{\beta}) = \beta$ as $N \to \infty$ and T is fixed
- we have asymptotic normality with fixed T and $N \to \infty$ if, in addition,
  - homoskedasticity: $var(u_{it}|\mathbf{X}_i, a_i) = \sigma^2$
  - no serial correlation: $cov(u_{it}, u_{is}|\mathbf{X}_i, a_i) = 0, \; t \neq s$
- estimation of the fixed-effects a_i is not consistent as $N \to \infty$ and T is fixed

FE Estimation: Final Remarks

- *FE* allows for arbitrary correlation between a_i and the vector of controls $\mathbf{X}$
- it is also called *WITHIN* estimator as it uses the time variation within each individual
- time-invariant controls dissapear in the transformation
- STATA does fixed effects as an option in *xtreg*

FD vs. *FE*

- *FD* and *FE* will give the same estimates when $T = 2$
- for $T > 2$, the two methods are different
- if u_{it} are uncorrelated, *FE* is more efficient than *FD*
- if Δu_{it} are uncorrelated, *FD* is better: test whether Δu_{it} are serially correlated
- always try both: if results are not very sensitive, good!

Random Effects

We can also impose more assumptions on $cov(a_i, X)$

$y_{it} = \beta_0 + \beta \mathbf{x}_{it} + a_i + u_{it}$

- if $cov(a_i, x_{it}) = 0 \Rightarrow \hat{\beta}_{\text{OLS}}$ will be consistent
- but not asym. efficient ($v_{it} = a_{iti} + u_{it}$ is serially correlated)
- we can do FGLS to improve efficiency

Random Effects

- RE is FGLS without serial correlation in u_{it}
- $y_{it} - \lambda \overline{y}_i = \beta(\mathbf{x}_{it} - \lambda \overline{\mathbf{x}}_i) + (v_{it} - \lambda \overline{v}_i)$
- if $cov(\mathbf{x}_{it}, u_{is}) = \mathbf{0}$ then $\hat{\beta}_{RE}$ will be consistent

RE and unobservable Heterogeneity

- if u_{it} is large relative to a_i, then $\hat{\lambda}$ will be close to 0 and *RE* will be similar to Pool OLS
- if u_{it} is small relative to a_i, then $\hat{\lambda}$ will be close to 1 and *RE* will be similar to *FE*
- STATA will do Random Effects for us as an option of *xtreg*

FE vs. *RE*

- if time-invariant unobserved heterogeneity is correlated with the controls, then *FE* is consistent while *RE* is not
- if random effects assumptions are true, *RE* will be more efficient than *FE*

The Haussman Test

- $H_0 : RE \Leftrightarrow H_0 : cov(a_i, \mathbf{x}_{it}) = 0$
- under the null, both *FE* and *RE* are consistent, but *RE* is asymptotically more efficient
- under the alternative, *FE* is still consistent
- the Hausman test compares the two estimators: a large difference suggests the null is false

Additional Issues

- you can test and correct for serial correlation and heteroskedasticity
- you can estimate standard errors which are robust to both
- it is possible to think of models where there is an unobserved fixed effect, even if we do not have the usual panel data structure (as in the twins example)
- if cluster effects are uncorrelated to controls, Pool OLS can be used, but the standard errors should be adjusted for cluster correlation

Summary

- Using cross-sections, we can test structural changes in the model
- A simple procedure available in pooled cross-sections to evaluate economic policies is the diff-in-diff estimator
- Panel data can be used to address some kinds of omitted variable bias
- *FE* is consistent under any correlation between time-invariant unobservable heterogeneity and the controls
- *RE* is the most efficient estimator in the absence of such correlation