Panel Data Analysis

Prf. José Fajardo
Fundação Getulio Vargas

Motivation

- More observations mean more information.
- More observations with a certain structure mean much more information: pooled cross sections and panel data
- How can we extract additional information from pooled cross sections or panel data?

Example

Example

Effect of an Incinerator on Housing Prices

With cross-sectional data in 1981, we have

$$\widehat{rprice} = 101,307.5 - 30,688.27\,nearinc$$
$$(3,093.0) \quad (5,827.71)$$
$$n = 142$$

With another cross-sectional data in 1978 when there were no incinerator, we have

$$\widehat{rprice} = 82,517.23 - 18,824.37\,nearinc$$
$$(2,653.79) \quad (4,744.59)$$
$$n = 179$$

Example cont'

Effect of an Incinerator on Housing Prices

With cross-sectional data in 1981, we have

$$\widehat{rprice} = 101,307.5 - 30,688.27\,nearinc$$
$$(3,093.0) \quad (5,827.71)$$
$$n = 142$$

With another cross-sectional data in 1978 when there were no incinerator, we have

$$\widehat{rprice} = 82,517.23 - 18,824.37\,nearinc$$
$$(2,653.79) \quad (4,744.59)$$
$$n = 179$$

Therefore, the true effect of the incinerator in not $-30,688.27$ but $-30,688.27 - (-18,824.37) = -11,863.90$.

Pooled cross section

- Pooled cross sections help us to evaluate the policy effect correctly by measuring the difference twice (before and after the policy implementation.)
- Recall the two regressions in the incinerator example:

$$rprice = \gamma_0 + \gamma_1 nearinc + u \quad \text{in years 1978 and 1981}$$
$$\hat{\delta}_1 = \hat{\gamma}_{1,81} - \hat{\gamma}_{1,78}$$
$$= \left(\overline{rprice}_{81,nr} - \overline{rprice}_{81,fr}\right) - \left(\overline{rprice}_{78,nr} - \overline{rprice}_{78,fr}\right)$$

- If perfectly randomized, the second term is 0.
- This estimator is called the Difference-in-Difference estimator.

Pooled cross section

- The effect can be estimated just by a single regression with some dummy variable.

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u$$

- This result is not intuitive. Just follow the logic:

	Before ($y81 = 0$)	After ($y81 = 1$)	After-Before
Control ($nearinc = 0$)	β_0	$\beta_0 + \delta_0$	δ_0
Treatment ($nearinc = 1$)	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment-Control	β_1	$\beta_1 + \delta_1$	δ_1

- Therefore, δ_1 in the above regression gives the same estimate of the Difference-in-Difference estimator.

Stata

```
use E:/pd/kielmc
reg rprice nearinc if year==1981
scalar b1=_b[nearinc]
reg rprice nearinc if year==1978
scalar b2=_b[nearinc]
display b1-b2
reg rprice nearinc y81 y81nrinc
reg rprice nearinc y81 y81nrinc age agesq
reg rprice nearinc y81 y81nrinc age agesq intst land area rooms baths
reg lprice nearinc y81 y81nrinc
reg lrprice nearinc y81 y81nrinc age agesq lintst lland larea rooms baths
reg lprice nearinc y81 y81nrinc age agesq lintst lland larea rooms baths
```

Data Structure

- A set of pooled cross sections is obtained by sampling randomly from a large population at different time points.
- A (typical) panel data set follow the same individuals over time.
- For example, consider that I sample three individuals from this room at two time points:

Time	Pooled	Panel
t=1	John, Jane, Evelyn	Eric, Andrew, Rachel
t=2	Kyle, Justin, Lisa	Eric, Andrew, Rachel

Data Structure

Table: Pooled Data

year	rprice	nearinc	y81
1978	60000	0	0
1978	54000	1	0
1978	38000	1	0
⋮	⋮	⋮	
1981	82000	1	1
1981	52000	0	1
1981	97000	0	1

Table: Panel Data

id	year	inf	unem
12	1950	7.3	3.5
12	1951	9.1	2.7
16	1950	5.3	5.4
16	1951	4.6	6.7
⋮	⋮	⋮	⋮
43	1950	7.1	4.2
43	1951	8.5	3.2
47	1950	6.7	5.4
47	1951	2.6	9.4

Data Structure

- There are also very useful panel structures other than the individual-time combination.
  1. Twins data: i is for twins id, and t is for the individual among the specific twins. Control for unobserved generic factors.
  2. School data: students sampled from many schools (or classrooms). Then, i is for school id, and t is for the student in school i.

What is Longitudinal data?

- Observed over time as well as over space.
- Pure cross-section data has many limitations. Problem is that only have one historical context.
- Time series allows for multiple historical context, but for only one spatial location.
- Longitudinal data - repeated observations on units observed over time

Types of Longitudinal data

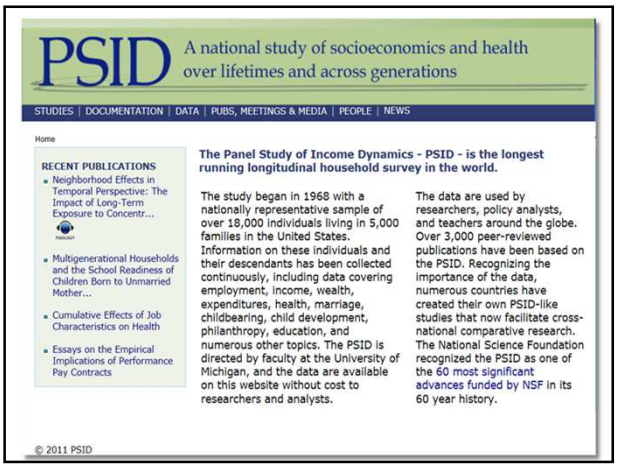
- "Panel study" (NES, PSID, Congressional election outcomes by CD and year)
- Often use panel data as a single "enriched" cross-section, with info on prior behavior
- "Time-Series–Cross-Section" (political economy data on 15 OECD nations observed annually)
- Data combining different surveys taken at different times
- Rolling Cross-Section (Canadian Election Study)
- "Pseudo Panel" (group respondents by cohort) based on "Repeated Cross Section Data" (eg Family Expenditure Surveys)

Panel vs TSCS data

- Logically TSCS data looks like panel data, but panels have large number of cross-sections (big *N*) with each unit observed only a few times (small *T*); TSCS data has reasonable sized *T* and not very large *N*. For panel data, asymptotics are in *N*, *T* is fixed. For TSCS data, asymptotics in T, N is fixed.
- This distinction is critical. Many of the panel methods are designed to deal with what is known as the "incidental parameters" problem, that is, as the number of parameters goes to ∞, one loses consistency. As we shall see this is a problem only for panel, not TSCS data.
- Furthermore, with small *T* there is no hope of saying anything about the time series structure of the data; with "biggish" *T* there is.
- We also care about the units in TSCS data; they are states or countries. We do not usually care about the units in panel models; they are just a sample, and we care about the population parameters, not the sample.

Specific examples - PSID

- Panel Study of Income Dynamics
- Based at SRC, University of Michigan
- Began in 1968 with 4,800 households.
- Original sample combined representative cross-section and low-income sample. Now has around 7,000 households.
- Annual interviews 1968-96, biennial since 1997, with household head (but covering all houseold members)
- Face-to-face PAPI 1968-72, mainly telephone interviewing (CATI) since 1973.
- http://psidonline.isr.umich.edu/

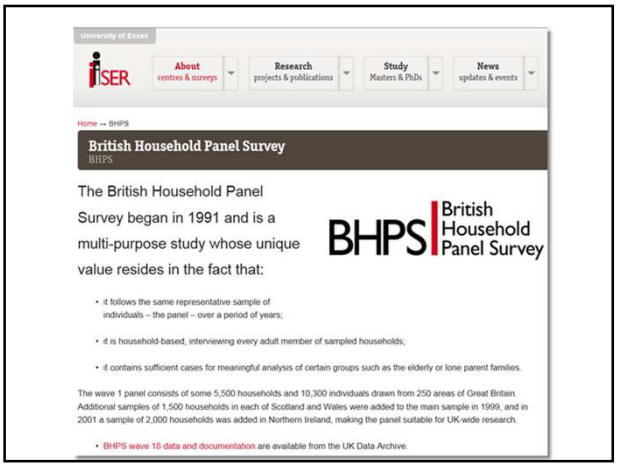


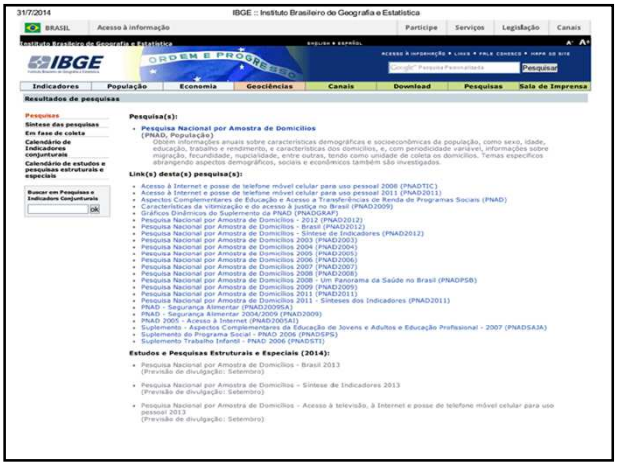
Specific examples - GSOEP

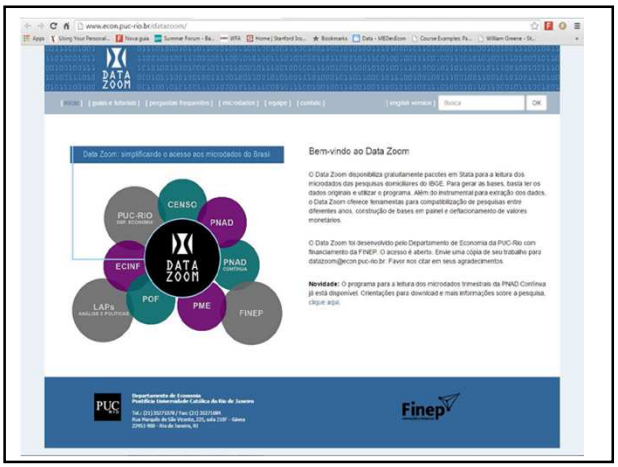
- German Socio-Economic Panel Study
- Based at DIW, Berlin
- Began in 1984 with approx 6,000 households.
- Various "top-ups" including expansion to former GDR. Now has around 12 000 households.
- Annual interviews with all adult members of hh.
- Various interview modes with gradual introduction of CAPI (computer-aided personal interviewing) since 1998.
- http://www.diw.de/english/soep/

Specific examples – BHPS/UKHLS

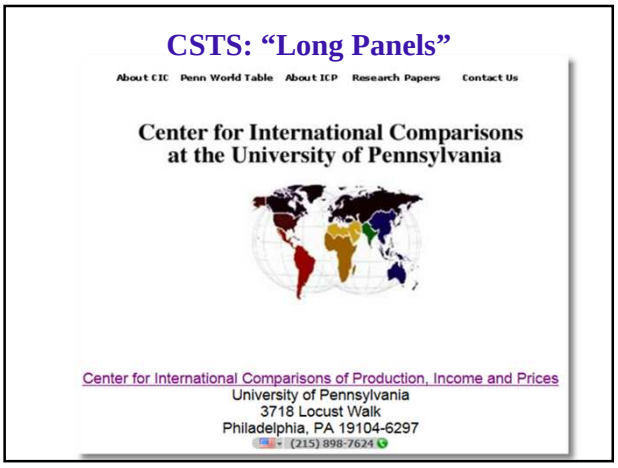
- British Household Panel Survey. Based at ISER, University of Essex
- Began in 1991 with approx 5,500 households (approx 10,000 adults) from England, Wales and (most of) Scotland. Extension samples from Scotland and Wales (1500 households each) added in 1999; sample from Northern Ireland (2000 households) added in 2001.
- Annual interviews with all adults (aged 16+) in household. Interviews with 11-16s added in 1994
- Questionnaires have annually-repeated core + less frequent or irregular additions. CAPI since 1999
- http://www.iser.essex.ac.uk/survey/bhps
- Now absorbed into the UK Household Longitudinal Survey (*Understanding Society*) with 40,000 households
- http://www.understandingsociety.org.uk/

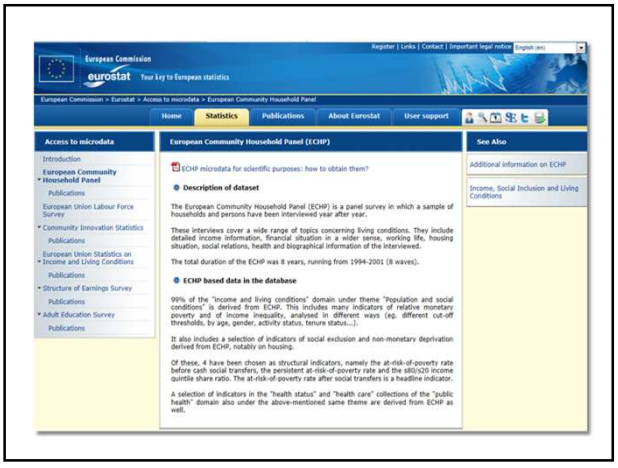






CSTS: "Long Panels"





"Huge Panels"

And more data...

- https://sites.google.com/site/medevecon/development-economics/devecondata

Model Building

José Fajardo
FGV/EBAPE

Objectives in Model Building

- **Specification**: guided by underlying theory
  - Modeling framework
  - Functional forms
- **Estimation**: coefficients, partial effects, model implications
- **Statistical inference**: hypothesis testing
- **Prediction**: individual and aggregate
- **Model assessment** (fit, adequacy) and evaluation
- **Model extensions**
  - Interdependencies, multiple part models
  - Heterogeneity
  - Endogeneity
- **Exploration**: Estimation and inference methods

Why use panel data?

- Can isolate effects of unobserved differences between individuals
- Causal inference may be strengthened by temporal ordering
- Repeated current observation more reliable than recall of histories in one-shot cross-section surveys
- Some phenomena are inherently longitudinal (e.g. poverty persistence; unstable employment)
- Can study dynamics – may be important even if we're only interested in the long run

Example: $y_{it} = \alpha x_{it} + \beta x_{it-1} + u_{it}$

$\Rightarrow$ Long-run impact = $\alpha + \beta$

Regression of y_{it} on $x_{it} \Rightarrow$

coefficient $b_{yx} = \alpha + \beta\, cov(x_{it}, x_{it-1})/var(x_{it}) \neq \alpha + \beta$

So a static model doesn't necessarily give good estimates of the long-run relationship

Limitations

BUT don't expect too much…

- Variation between people usually far exceeds variation over time for an individual

  $\Rightarrow$ a panel with T waves doesn't give T times the information of a cross-section
- Variation over time may not exist for some important variables
- Variation over time may be inflated by measurement error
- Panel data imposes a fixed timing structure; continuous-time survival analysis may be more informative
- We still need very strong assumptions to draw clear inferences from panels: sequencing in time does *not* necessarily reflect causation

Some terminology

A **balanced panel** has the same number of time observations (T) for each of the n individuals

An **unbalanced panel** has different numbers of time observations (T_i) on each individual

A **compact panel** covers only consecutive time periods for each individual – there are no "gaps"

Attrition is the process of drop-out of individuals from the panel, leading to an unbalanced and possibly non-compact panel

A **short panel** has a large number of individuals but few time observations on each, (e.g. HILDA has 7,000 households and 10 waves)

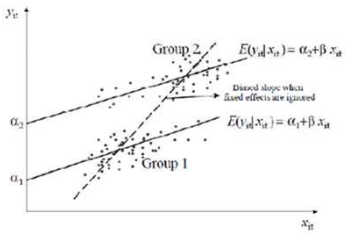
A **long panel** has a long run of time observations on each individual, permitting separate time-series analysis for each

Econometric Analysis of Panel Data

- Introduction
  - Panel Data Definition
    - Unbalanced Panel $y_{it}, \mathbf{x}_{it}\ (t = 1, 2, ..., T_i; i = 1, ..., N)$
    - Balanced Panel: $T_i = T, \forall i$
    - Short Panel: $T < \infty, N \to \infty$
    - Long Panel: $T \to \infty, N < \infty$
  - Panel Data Analysis
    - Unobserved Heterogeneity
    - Cross Section and Time Series Correlation

$$y_{it} = \mathbf{x}_{it}\beta + u_i + e_{it}$$

Bias from Ignoring Fixed Effects



Panel Data with Two Periods

- It's possible to use a panel just like pooled cross-sections, but can do more than that
- Panel data can be used to address some kinds of omitted variable bias
- If can think of the omitted variables as being fixed over time, then can model as having a composite error

34

Panel Data with Two Periods

- Suppose the population model is

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$$

- Here we have added a time-constant component to the error, $u_{it} = a_i + u_{it}$
- If a_i is correlated with the x's, OLS will be biased, since we a_i is part of the error term
- With panel data, we can difference-out the unobserved fixed effect

35

First Differences

- We can subtract one period from the other, to obtain $\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + \ldots + \beta_k \Delta x_{ik} + \Delta u_i$
- This model has no correlation between the x's and the error term, so no bias
- Need to be careful about organization of the data to be sure compute correct change

36

Example

```
use e:/pd/crime2
reg crmrte unem if year==87
reg crmrte unem d87
reg ccrmrte cunem
estat hettest
estat ovtest
```

Panel data Analysis Using Stata

- Declare panel data and variables
  - xtset
- Panel data analysis: xt commands
  - xtdes
  - xtsum
  - xtdata
  - xtline
- Panel data regression
  - xtreg

Panel data Analysis Using Stata

- Hypothesis Testing
  - xthausman
  - xttest0
- Advanced Topics
  - xtregar
  - xthtaylor (Hausman-Taylor Estimator)
  - xtivreg (Instrumental Variables Estimation)
  - xtabond (Arellano-Bond Estimator)

Example: Returns to Schooling

•Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators"

Journal of Applied Econometrics, 3, 1988, pp. 149-155.

Data Source: Panel Study of Income Dynamics:

595 Individuals, 7 Years

Example: Returns to Schooling

- LWAGE = log of wage
- EXP = work experience
  WKS = weeks worked
  OCC = occupation, 1 if blue collar,
  IND = 1 if manufacturing industry
  SOUTH = 1 if resides in south
  SMSA = 1 if resides in a city (SMSA)
  MS = 1 if married
  FEM = 1 if female
  UNION = 1 if wage set by union contract
  ED = years of education
  BLK = 1 if individual is black

Example: Returns to Schooling

The Model:

$$LWAGE_{it} = \beta_0 + \beta_1 EXP_{it} + \beta_2 EXP_{it}^2 + \beta_3 WKS_{it} + \beta_4 OCC_{it} + \beta_5 IND_{it} + \beta_6 SOUTH + \beta_7 SMSA + \beta_8 MS_{it} + \beta_9 UNION_{it} + \beta_{10} ED_i + \beta_{11} FEM_i + \beta_{12} BLK_i + \varepsilon_{it}$$

Stata

```
clear
*input data
insheet using e:\pd\TableF8-1.csv
% infile exp wks occ ind south smsa ms fem union ed blk lwage using e:\paneldata\TableF8-1.csv
describe
summarize
generate person=group(595)
bysort person: generate period=group(7)
* panel data definition
xtset person period
xtdes
xtsum
*one-way tabulation of data
xttab union
xttab ind
xttrans ms
xttab ed // ed is time invariant
```

Stata

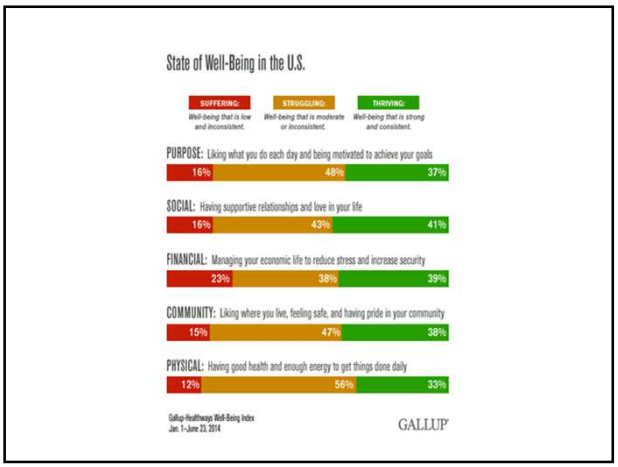
```
*plots of panel data
xtline lwage if person<=10, overlay
graph twoway (scatter lwage exp)
generate exp2=exp^2
local x1 exp exp2 wks occ ind south smsa ms union
local x2 ed blk fem
* panel data regression: y=lwage
*x1=[1 exp exp2 wks occ ind south smsa ms union],
*x2=[ed blk fem] (time-invariant regressors)
regress lwage `x1' `x2'
regress lwage `x1' `x2', vce(cluster person)
```

Reading

- Marriage Premium
- Does marriage make people happy, or do happy people get married?
- Well-Being and Happiness

  "How satisfied are you with your life as a whole?"

"We draw pleasure and pain from what is happening at the moment, if we attend to it." Daniel Kahneman in *Thinking, Fast and Slow*.



Percentage of Americans Thriving in Each Element of Well-Being

AGE	Purpose	Social	Financial	Community	Physical
18-29 years	38	38	34	30	36
30-44 years	35	36	30	34	31
45-64 years	33	38	35	37	28
65+	44	53	62	53	40

GENDER	Purpose	Social	Financial	Community	Physical
Men	33	39	39	36	30
Women	40	42	39	40	35

REGION	Purpose	Social	Financial	Community	Physical
East	34	40	38	35	33
Midwest	35	38	40	36	31
South	39	42	38	40	32
West	37	41	40	39	35

Gallup-Healthways Well-Being Index
Jan. 1-June 23, 2014
GALLUP