EC968

Panel Data Analysis

Steve Pudney

ISER

University of Essex

ISER

Course structure

Lecture 1: Basics

- Basic concepts
- Summarising panel data
- Example: BHPS wages data
- Unobservables & identification of their effects

Lecture 2: Linear regression for panel data

- Within-group ("fixed effects")regression
- Asymptotics for short panels
- Random effects regression
- Testing the zero covariance assumption

Lecture 3: Instrumental variable estimation

- Correlated individual effects: Hausman-Taylor estimation
- Endogenous regressors: the within-group IV estimator
- Dynamic regression models

Lecture 4: Discrete models

- Binary variables: conditional logit
- Random effects models with state dependence

Seminar: Stata applications

- Group mini-projects

Lecture 1: Basics

- Basic concepts
- Summarising panel data
- Example: BHPS wages data
- Unobservables & identification of their effects

What are Panel Data?

Panel data are a form of longitudinal data, involving regularly repeated observations on the same individuals

Individuals may be people, households, firms, areas, etc

Repeat observations may be different time periods or units within clusters (e.g. workers within firms; siblings within twin pairs)

Some types of panel data

- **Cohort surveys**
  - Birth cohorts (NCDS, British Cohort Survey 1970, Millennium CS)
  - Age group cohorts (NLSY, MtF, Addhealth, HRS, ELSA)
  - Many programme evaluation studies and social experiments
- **Panel surveys**
  - Rotating household panels: (Labour Force Surveys, US SIPP)
  - Perpetual household panels: an indefinitely long horizon of regular repeated measurements
  - Company panels: firms observed over time, linked to annual accounts information
- **Non-temporal survey panels**
  - Example: Workplace Employment Relations Survey (WERS) ⇒ cross-section of workplaces, 25 workers sampled within each
- **Non-survey panels** (aggregate panels)
  - countries, regions, industries, etc. observed over time
- **Useful catalogue** of longitudinal data resources:

  http://www.iser.essex.ac.uk/ulsc/keeptrack/index.php

The BHPS

http://www.iser.essex.ac.uk/ulsc/bhps/

- British Household Panel Survey, based at ISER, University of Essex
- Began in 1991 with approx 5,500 households (approx 10,000 adults)
- 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.
- Youth and child interviews added in 1994 & 2002
- Questionnaires have annually-repeated core + less frequent or irregular additions
- Now CAPI
- See BHPS quality profile for technical detail (http://www.iser.essex.ac.uk/ulsc/bhps/quality-profiles/BHPS-QP-01-03-06-v2.pdf)

Some terminology

A **balanced panel** has the same number of time observations (T) on 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. BHPS has 5,500 households and 13 waves)

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

We consider mainly short panels in this course

Basic notation

We work with observed variables y_{it} , $\mathbf{z}_i$ and $\mathbf{x}_{it}$, where:

y_{it} = dependent variable to be analysed

$\mathbf{z}_i$ = row-vector of k_z time-invariant characteristics
(*e.g.* year of birth, sex)

$\mathbf{x}_{it}$ = row-vector of k_x time-varying characteristics
(*e.g.* job tenure, marital status)

where i indexes individuals, t indexes time periods.

y_{it} may be:

- continuous (*e.g.* wages);
- mixed discrete/continuous (*e.g.* hours of work);
- binary (*e.g.* employed/not employed);
- ordered discrete (*e.g.* Likert scale for degree of happiness);
- unordered discrete (*e.g.* occupation)

Disadvantages of cross-section data

Example: cross-section Mincer earnings equation (t subscript suppressed)

$$y_i = \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_{it}$$

where:

y_i = log wage;

$\mathbf{z}_i$ = observable time-invariant factors (education, etc.);

$\mathbf{x}_i$ = observable time-varying factors (*e.g.* job tenure);

ε_i = random error (*e.g.* "luck")

Possible misspecifications, causing bias:

- Omitted dynamics (lagged variables not observed)
- Reverse causation (*e.g.* pay and tenure jointly determined)
- Omitted unobservables (*e.g.* "ability")

Advantages of panel data

With panel data:

- We can study dynamics
- The sequence of events in time helps to reveal causation
- We can allow for time-invariant unobservable variables

BUT…

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

  ⇒ a panel with *T* waves doesn't give *T* times the information of a cross-section
- Variation over time may not exist or may be inflated by measurement error
- Panel data imposes a fixed timing structure; continuous-time survival analysis may be more informative

Summarising panel data

There are various sensible ways to get a general idea of the nature of your data. For example:

- Between- and within-group components of variation
- Cohort profiles
- Transition tables

Important Stata commands:

tsset - defines variables to identify *i* and *t* for each case

xtdes – describes the pattern of available cases

xtsum – gives between & within-group decomposition

xttrans – calculates transition matrices

(but note: care needed for non-compact panels)

Sample Stata programme in downloadable file EC968earnings.do

Between- and within-group variation

Define the individual-specific or group mean for any variable, *e.g.* y_{it} as:

$$\bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it}$$

y_{it} can be decomposed into 2 orthogonal components:

$$y_{it} - \bar{y} = (y_{it} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

$$= \text{within} \quad + \quad \text{between}$$

where $\bar{y} = \sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{it} \Big/ \sum_{i=1}^{n} T_i$

Corresponding decomposition of sum of squares:

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it} - \bar{y})^2 = \sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it} - \bar{y}_i)^2 + \sum_{i=1}^{n}\sum_{t=1}^{T_i}(\bar{y}_i - \bar{y})^2$$

or: $T_{yy} = W_{yy} + B_{yy}$

BHPS example: group-wise decomposition of earnings data

Sample of adult males and females with earnings & hours data

y_{it} = hourly wage (2000 prices); mean = 9.39

n = 5,860;

$\max(T_i)$ = 11; $\bar{T}$ = 3.6

Total sample size = $n\bar{T} = \sum T_i$ = 21,125

Total root mean square = 6.323

Within-group root mean square = 2.660

Between-group root mean square = 5.777

⇒ approx. 80% of the sample variance of wages is between-individual

Warning: measurement error may induce spurious variation

```
. xtsum          /*  Note measurement error in birth cohort variable !!!!! */

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
cohort

         overall |  1959.083   10.35489       1931       1980 |     N =   21124
         between |             11.4243        1931       1980 |     n =    5859
         within  |             .0133368   1958.483    1959.94 | T-bar = 3.60539
```

Transitions

- Want to compare state in this wave with state in last wave. Example: part-time work status (binary variable PT)
- If we have `tsset` the data, can easily create lagged values of variable: `generate lpt = l.pt`
- Then tabulate current against lagged value: `tabulate lpt pt`

```
. tabulate lpt pt, row

           | Part-time (<=30 hours
 Lagged PT |        total)
      work |         0          1 |     Total
-----------+----------------------+----------
         0 |    10,619        310 |    10,929
           |     97.16       2.84 |    100.00
-----------+----------------------+----------
         1 |       333      2,166 |     2,499
           |     13.33      86.67 |    100.00
-----------+----------------------+----------
     Total |    10,952      2,476 |    13,428
           |     81.56      18.44 |    100.00
```

- Same result with command: `xttrans pt, freq`

Transitions and measurement error

Analysis of transitions can give good indications of data (un)reliability
Example: UK Offending Crime & Justice Survey (2003-4, ages 10-25)

```
. xttrans d1evec, freq
```

have you ever taken cannabis	have you ever taken cannabis Yes	No	DK	DWTA	Total
Yes	728 86.67	111 13.21	0 0.00	1 0.12	840 100.00
No	251 10.23	2,189 89.24	6 0.24	7 0.29	2,453 100.00
DK	2 15.38	9 69.23	1 7.69	1 7.69	13 100.00
DWTA	9 60.00	5 33.33	0 0.00	1 6.67	15 100.00
Total	990 29.81	2,314 69.68	7 0.21	10 0.30	3,321 100.00

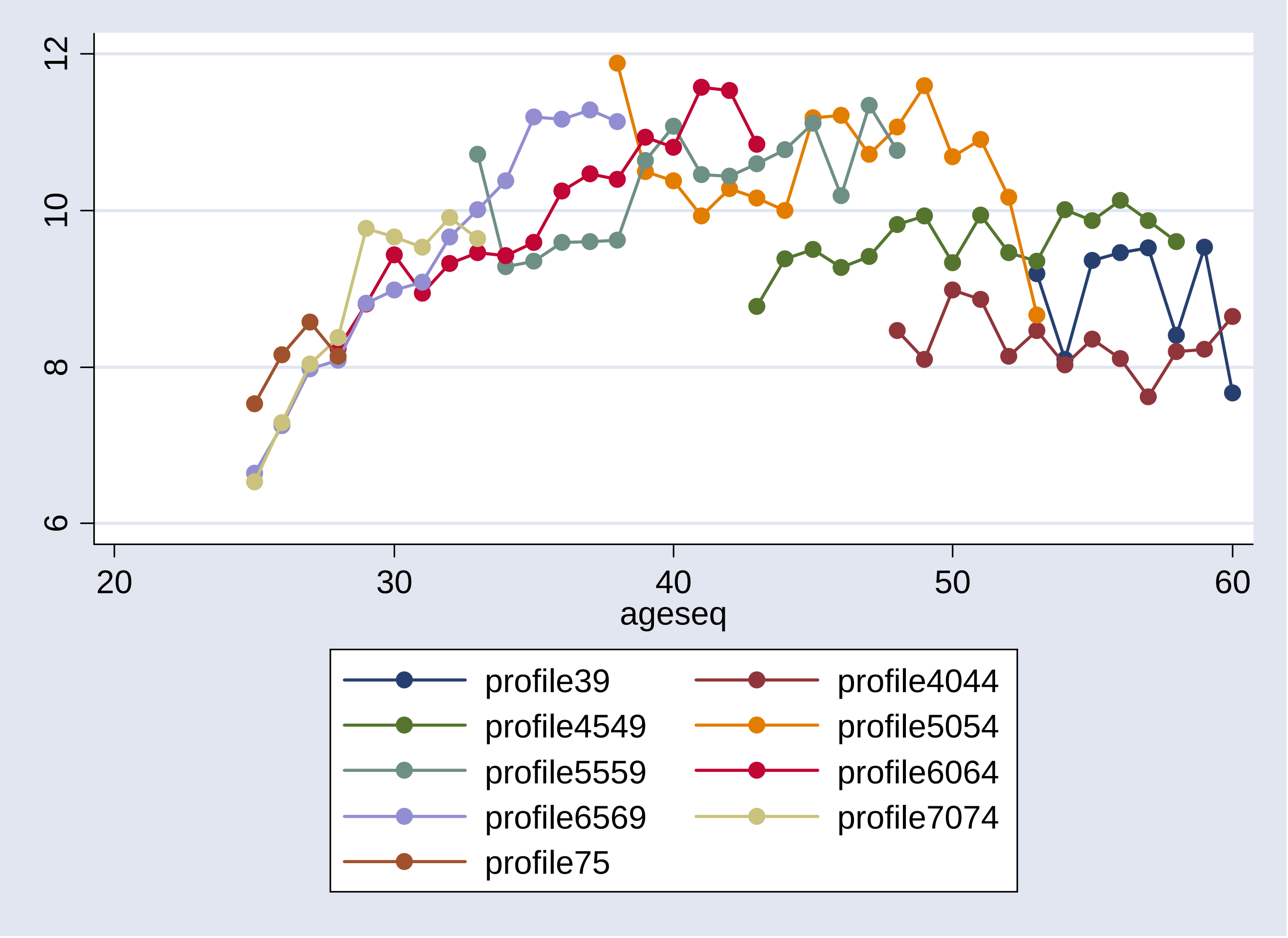
13% of people who'd used cannabis before 2003 say they've never used before 2004!!

BHPS example: earnings transition rates

Pay groups: 1 = under £5.00; 2 = £5-7; 3 = £7-10; 4 = £10-15; 5 = £15 and over

	lagpaygrp					
paygrp	1	2	3	4	5	Total
1	1,443 68.26	381 16.87	67 2.54	21 0.85	3 0.17	1,915 16.95
2	550 26.02	1,246 55.16	370 14.03	30 1.22	4 0.22	2,200 19.48
3	95 4.49	569 25.19	1,604 60.83	324 13.12	23 1.27	2,615 23.15
4	22 1.04	54 2.39	563 21.35	1,694 68.61	211 11.61	2,544 22.52
5	4 0.19	9 0.40	33 1.25	400 16.20	1,576 86.74	2,022 17.90
Total	2,114 100.00	2,259 100.00	2,637 100.00	2,469 100.00	1,817 100.00	11,296 100.00

BHPS example: cohort earnings profiles



Two basic identification problems

(1) Unobservable variables

- Can we distinguish the impact of unobservables from general serial correlation?
- Can we distinguish the impact of unobservables from the impact of time-invariant observables?

(2) Age, cohort and time effects – can they be distinguished?

- Behaviour may change with age
- Current behaviour may be affected by experience in "formative years" ⇒ cohort or year-of-birth effect
- Time may affect behaviour through changing macro environment

Identification problem (1): Unobservables

Example: Mincer wage models based on human capital theory:

$$y_{it} = \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

where:

y_{it} = log wage

$\mathbf{z}_i$ = observable time-invariant factors (*e.g.* education)

$\mathbf{x}_{it}$ = observable time-varying factors (*e.g.* job tenure)

u_i = unobservable "ability" (assumed not to change over time)

ε_{it} = "luck"

Pooled data regression of y on $\mathbf{z}$ and $\mathbf{x} \Rightarrow$ omitted variable bias:

$$\text{bias}\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \text{regression of } u \text{ on } (\mathbf{z}, \mathbf{x})$$

Ability (u) is likely to be positively related to education ($\mathbf{x}$)

$\Rightarrow$ bias in estimate of returns to education

Unobservables: the identification problem

It may seem puzzling that panel data allows us to draw conclusions about the impact of a variable without observing it
– and there are good reasons for being puzzled!

Consider the *nonparametric* identification problem:

Define $\mathbf{y}_i = (y_{i1} \ldots y_{iT})$; $\mathbf{X}_i = (\mathbf{x}_{i1} \ldots \mathbf{x}_{iT})$.

If we know the distribution of $\mathbf{y}_i \mid (\mathbf{z}_i , \mathbf{X}_i)$ from sample data, can we infer the distribution of $\mathbf{y}_i \mid (\mathbf{z}_i , \mathbf{X}_i , u_i)$ without making assumptions about the distribution of u_i and its correlation with $\mathbf{z}_i$ and $\mathbf{X}_i$?

In general, the answer to this is obviously "no"…

Identification when all covariates are time-varying

E.g., assume a simple case:

- all covariates are time-varying, so there is no z_i in the model
- y_{it} can take q discrete values, so $\mathbf{y}_i$ can take q^T possible values, $\Rightarrow q^T$-1 probabilities to be determined
- $\mathbf{x}_{it}$ is a single categorical variable, taking r possible values; thus $\mathbf{X}_i$ can take r^T possible values

So $\mathbf{y}_i \mid \mathbf{X}_i$ is a set of r^T distributions, each with q^T-1 probabilities. So there are $(q^T\text{-}1) \times r^T$ known items of information. From this, we want to infer the distributions $\mathbf{y}_i \mid (\mathbf{X}_i, u_i)$, containing $(q^T\text{-}1) \times r^T \times s$ probabilities, where s is the number of possible ability levels.

Therefore, we have more unknowns than knowns whenever $s > 1$ (*i.e.* if ability varies)

$\Rightarrow$ the distribution of y conditional on (x, u) is not identified.

Identifying assumptions

- To solve the identification problem, we make strong assumptions, particularly: ***conditional serial independence***
- Assume y_{i1} ... y_{iT} are known *a priori* to be independent, conditional on ($\mathbf{X}_i$, u_i). Rather than q^T-1 probabilities to be determined given ($\mathbf{X}_i$, u_i), there are only T(q-1) (*i.e. q-1* probabilities for each period). This implies T (q-1)r^Ts probabilities for $\mathbf{y}_i$ | ($\mathbf{X}_i$, u_i) to be determined from the (q-1)$^Tr^T$ known probabilities of $\mathbf{y}_i$ | $\mathbf{X}_i$. A necessary condition is that the number of knowns exceeds the number of unknowns: $(q^T\text{-}1)\ r^T \geq T(q\text{-}1)\ r^Ts$

or: $$(q^T\text{-}1)\ /\ T(q\text{-}1)\ \geq\ s$$

- This is satisfied when T and q are sufficiently large, relative to s. *I.e.*, detailed identification is possible if y is sufficiently close to continuous variation and if the panel is sufficiently long.
- *E.g*. if y is binary (q = 2), T = 4 waves will only identify s = 3 ability levels; if y can take 3 values, 4 waves will identify 10 ability levels.

Identification with time-invariant covariates: can we distinguish $\mathbf{z}_i$ and u_i?

Consider the distribution $f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, u_i)$. Let $h(u_i, \mathbf{z}_i)$ be an arbitrary function, invertible with respect to u_i and construct a new unobservable:

$$v_i = h(u_i, \mathbf{z}_i)$$

Then:

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, u_i) \equiv f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, h^{-1}(v_i, \mathbf{z}_i))$$

Call the right-hand side of this $g(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, v_i)$. Then:

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, u_i) \equiv g(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{z}_i, v_i)$$

Therefore, the functions $f(.)$ and $g(.)$ are equally valid descriptions of the data. They involve the same observable variables but different unobservables.

So the distribution $f(.)$ is not identifiable without further restrictions. For example, we could assume that u_i and $\mathbf{z}_i$ are independent. That would rule out $v_i = h(u_i, \mathbf{z}_i)$ as a valid unobservable, since $h(u_i, \mathbf{z}_i)$ is not independent of $\mathbf{z}_i$.

Implications

In models like:

$$y_{it} = \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

- We can only identify the effect of unobservable ability u_i if we can assume that ε_{it} is serially-independent (or has a highly restricted autocorrelation structure).
- We cannot distinguish the separate effects of $\mathbf{z}_i$ and u_i without making further assumptions.

Identification problem (2): Age, cohort & time effects

Fundamental identity relating age (A_{it}), time of interview (t) and birth cohort (B_i):

$$A_{it} \equiv t - B_i$$

These three cannot be distinguished in principle. To do so would require an ability to move a cohort forward or back in time (!) to measure the effect of time holding age and cohort constant.

- In a cross-section, t doesn't vary, so time effects can't be estimated and age or cohort are collinear – only their joint effect can be estimated
- In a panel, two of the three effects can be estimated. *E.g.* the following model can be rewritten in several equivalent ways

$$\begin{aligned} y_{it} &= h(A_{it}, t, B_i) + u_i + \varepsilon_{it} \\ &= h(t - B_i, t, B_i) + u_i + \varepsilon_{it} \equiv h_2(t, B_i) + u_i + \varepsilon_{it} \\ &= h(A_{it}, A_{it}+B_i, B_i) + u_i + \varepsilon_{it} \equiv h_3(A_{it}, B_i) + u_i + \varepsilon_{it} \\ &= h(A_{it}, t, t - A_{it}) + u_i + \varepsilon_{it} \equiv h_4(A_{it}, t) + u_i + \varepsilon_{it} \end{aligned}$$

So we can use (t, B_i), (A_{it}, B_i) or (A_{it}, t) as covariates, but not all three.

Age, cohort and time effects

A possible solution is to think more deeply about the effects of time and cohort and introduce further information.

E.g. we may think it is current macro-level conditions at the time of birth that generate differences between cohorts and current macro conditions that generate time effects.

Let $\mathbf{w}(t)$ be the vector of relevant macro variables at historical time t.

Then our model would be:

$$\begin{aligned} y_{it} &= h(A_{it}, t, B_i) + u_i + \varepsilon_{it} \\ &= h(A_{it}, \mathbf{w}(t), \mathbf{w}(B_i)) + u_i + \varepsilon_{it} \end{aligned}$$

This breaks the exact functional relationship between age, time and cohort effects and permits identification.