Lectures on:
Panel data analysis for social scientists, given at the University of Bergen, October 2006

You may find these lecture notes a useful complement to those I will use for EC968. They cover a wider range of topics and go at a slower pace, with less emphasis on technical issues.

Steve Pudney

Panel Data Analysis for Social Scientists

University of Bergen

Department of Sociology

Department of Comparative Politics

Steve Pudney

Gabriella Conti

ISER

Aims of course

- Introduce the distinctive features of panel data.
- Review some panel data sets commonly used in social sciences.
- Present the advantages (and limitations) of panel data, and consider what sort of questions panel data can(not) address.
- Show how to handle and describe panel data.
- Introduce the basic estimation techniques for panel data
- Discuss how to choose (and test for) the right technique for the question being addressed.
- Discuss interpretation of results

Structure of course (1)

- 5 days × (3 hours lectures + 2 hour lab sessions)
- Lab sessions will illustrate concepts using Stata software ("industry standard" in survey-based applied work)
- Main data will be from British Household Panel Survey (BHPS)
- Focus is on understanding the concepts and applying them.
- Full lecture slides on the web
- Technical detail kept to a minimum but available in "appendices"

Structure of course (2)

Day 1: Basics

- What are panel data (examples)?
- Why use panel data?
- Handling panel data in Stata – some basic commands.
- Patterns of observations in panel data (non-response and attrition)
- Within and between variation
- Transitions.
- Cohort analysis

Day 2: Statistical analysis

- Inference using panel data: some identification issues
  - unobservables.
  - age, time and cohort effects
- Regression analysis: Within and between group regression

Structure of course (3)

Day 3: Random effects and endogeneity

- Random effects regression
- Testing the FE and RE assumptions
  - Hausman test
  - Mundlak model
- Endogeneity
  - The source of endogeneity
  - The between- and within-group IV estimator
  - Correlated individual effects: Hausman-Taylor estimation

Structure of course (4)

Day 4: Binary response models

- Types of discrete variables
- Why not linear regression?
- Latent linear regression
- Conditional (fixed-effects) logit
- Static random effects logit and probit
- Ordered response models

Day 5: Further topics

- Incomplete panels and sample selection in panel data models
- Dynamic fixed-effects models
- Count data models
- Policy evaluation and panel data

Day 1: Basics

- What are panel data
- Why use panel data?
- Handling panel data in Stata

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

Long-term household panels

- Individuals in their household context
- Perpetual panel survey, often with retrospective elements (period before first wave; periods between waves)
- Designed to maintain representativeness of the sampled population over time
- But may use refreshment samples if, e.g., substantial immigration, worries about panel fatigue/conditioning
- Examples worldwide, include
  - US PSID, Dutch HP, Swedish LoLS, German SOEP, BHPS, Canadian SLID, Australian HILDA, NZ SoFIE, European Community Household Panel, BHePS, NHPS, and several in developing countries (e.g. Indonesia, Ethiopia, VietNam)
- Big differences in: content, following rules, who is interviewed, interview method, etc.

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. Almost no phone interviews.

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)

Using household panels (1)

- Panel data involve regularly *repeated observations* on the same *individuals*.
- In most analysis using household panels, the *individual* is the **person** and the *repeated observations* are the **different time periods (waves)**. This is the case we will mostly consider.
- Sometimes, e.g. to isolate household (or family) effects, the *individual* is the **household** (or family) and the *repeated observations* are **different persons** within the household
- Multi-level analysis involves more than 2 dimensions of the sample, e.g. time periods within persons within households

Using household panels (2)

- Conceptual problems with **households** over **successive time periods (waves)**
  - households change their composition over time
  - how much can a hh change before it is effectively a new household?.
- We usually follow **persons** over **time periods (waves)** and treat household data as contextual information
  - e.g. an individual's material living standards measured as the income of their household at that time.
  - Rationale for household panel designs, rather than simpler cohort designs
  - Allows for individuals moving between households & forming new households

Why use panel data?

- Repeated observations on individuals allow for possibility of isolating effects of unobserved differences between individuals
- We can study dynamics
- The ability to make causal inference is enhanced by temporal ordering
- Some phenomena are inherently longitudinal (e.g. poverty persistence; unstable employment)
- Net versus gross change: gross change visible only from longitudinal data, e.g. decomposition of change in unemployment rate over time into contributions from inflows and outflows

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 or 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. BHPS has 5,500 households and 14 waves)

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

We consider only short panels in this course

Handling panel data in Stata

- For our purposes, the *unit of analysis* or *case* is either the person or household:
  - If case = person, case contains information on person's state, perhaps at different dates
  - If case = household, case contains info on some or all household members (cross-sectional only!)
- The data can be organised in two ways:
  - Wide form – data is sometimes supplied in this format
  - Long form – usually most convenient & needed for most panel data commands in Stata
  - Use Stata `reshape` command to convert between them.
- Three important operations:
  - Matching/merging
  - Aggregating
  - Appending

Wide format

- One row per case
- Observations on a variable for different time periods (or dates) held in different columns
- Variable name identifies time (via prefix)

PID	awage (Wage at w1)	bwage (Wage at w2)	cwage (Wage at w3)
10001	7.2	7.5	7.7
10002	6.3	missing	6.3
10003	5.4	5.4	missing
…			

Long format

- potentially multiple rows per case, with
- observations on a variable for different time periods (or dates) held in extra rows for each individual
- case-row identifier identifies time (e.g. PID, wave)

PID	wave	wage
10001	1	7.2
10001	2	7.5
10001	3	7.7
10002	1	6.3
10002	3	6.3
10003	1	5.4
10003	2	5.4
…	…	

Matching (or merging)

- Joining two (or more) files at the same level of observation (e.g. person files) where both (all) files contain the same *identifier* variable used as key
- 1:1 matching – one case in "master file" corresponds to one case in "using file" (i.e. the file being matched in)
- 1:many – one case in the "using file" may be 'distributed' to many cases in the "master file"
  - E.g. info. about a household attached to each one of the household's members
- In either case, not all cases in master file may receive match; not all cases in the using file may provide a match
- Stata's command: `merge` *`key`* `using` *`file`*
  - Merging is the source of many disastrous errors – always check by using `tabulate _merge` (see examples later)

Aggregation

- Deriving group-level information from all the members of that group
  - E.g. calculating household income from the incomes of its members
  - E.g. calculating how many children a woman has during her first marriage
- The group-level information may be used in two ways:
  - (i) saved in a new file with the group – e.g. household or spell – as the case (`collapse`)
  - (ii) attributed to each of the group members within the existing file (`egen`; `by(sort): ...`)

Appending

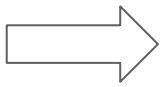
- Combining files with no index-based matching
  - E.g. combining file A with *n1* rows and file B with *n2* rows to produce a new file C with *n1*+*n2* rows.
- Stata command: `append`
- Used to assemble a sequence of annual cross-section data files into a single long-format panel data file
  - Rows in new combined files are specific to a person-wave combination
- Each variable must have the same name in each of the annual cross-section files

Sorting (ordering) the data

- We now have a dataset in long format
- It's a good idea to order the data for easier viewing. "Eyeballing" the data is important!
- We also have to tell Stata which variable identifies the *individual* (Stata calls this the panel variable).
- We may also have to tell Stata which variable identifies the *repeated observation* (Stata calls this the time variable).
  - For some types of panel analysis we don't need to know the ordering of the repeated observations

```
sort pid wave
```

PID	wave	wage
10001	2	7.5
10002	3	6.3
10002	1	6.3
10001	1	7.2
10001	3	7.7
10003	1	5.4
10003	2	5.4
…	…	



PID	wave	wage
10001	1	7.2
10001	2	7.5
10001	3	7.7
10002	1	6.3
10002	3	6.3
10003	1	5.4
10003	2	5.4
…	…	

***Note*: this panel is neither balanced nor compact**

Panel and time variables

- Use `tsset` to tell Stata which are panel and time variables:

```
. tsset pid wave
panel variable:  pid, 10002251 to 1.347e+08
time variable:  wave, 1 to 13, but with gaps
```

- Note that `tsset` automatically sorts the data accordingly.

Describing panel data

- Ways of describing/summarising panel data:
  - Basic patterns of available cases
  - Between- and within-group components of variation
  - Transition tables
- Some basic notation:

  y_{it} is the "dependent variable" to be analysed
  - i indexes the individual (pid), $i = 1, 2, \ldots, n$
  - t indexes the repeated observation / time period (wave), $t = 1, 2, \ldots, T_i$
- 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)

Describe patterns of panel data: `xtdes`

```
. xtdes

pid:  10002251, 10004491, ..., 1.347e+08            n =      16082
wave:  1, 2, ..., 13                                 T =         13

           Delta(wave) = 1; (13-1)+1 = 13
           (pid*wave uniquely identifies each observation)

Distribution of T_i:   min    5%     25%     50%     75%     95%     max
                         1     1       2       7      13      13      13

Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------------
    4648     28.90   28.90 |  1111111111111
     997      6.20   35.10 |  1...........
     646      4.02   39.12 |  11..........
     376      2.34   41.46 |  ............1
     342      2.13   43.58 |  111.........
     327      2.03   45.62 |  1111.........
     261      1.62   47.24 |  ...........11
     254      1.58   48.82 |  .1..........
     251      1.56   50.38 |  ..........111
    7980     49.62  100.00 | (other patterns)
 ---------------------------+---------------
   16082    100.00         |  XXXXXXXXXXXXX
```

Between- and within-group variation (1)

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 components:

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

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

where $\bar{y} = \sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{it} \Big/ n\bar{T}$ and $\bar{T}$ is average no. of periods per case

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:

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

Between- and within-group variation (2)

- Between and within variation is the basis of linear panel regression. Important concept to understand.
- Simple example: balanced panel (n=1119, T = 13) of workers who have reported their wages.
- From `summarize`, we have grand mean wage ($\bar{y}$) = £9.84 per hour, and (overall) variance of wages = 32.63. Recall the standard formula for variance:

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \bar{y})^2}{n\bar{T} - 1} \equiv \frac{T_{yy}}{n\bar{T} - 1}$$

Between- and within-group variation (3)

- So T_{yy} is the variance multiplied by its *degrees of freedom* $n\bar{T} - 1$ = 1119*13 – 1 = 14546 (or can calculate T_{yy} 'by hand' in Stata – see example in computer lab).
- We get T_{yy} = 32.627956 * 14546 = 474606.3
- Can calculate B_{yy} and W_{yy} manually in Stata (see example in computer lab). We get:
  - B_{yy} = 358920.7
  - W_{yy} = 115685.6
  - Check that $B_{yy} + W_{yy} = T_{yy}$!!
- Proportion of between variation is B_{yy} / T_{yy} = 76%. Most variation is between people not within people! Measurement error may make this an *under*estimate!

Within and between deviations in the data

pid	wave	Wage	Grand mean	Ind. Mean	Within dev	Between dev	Total dev
10028005	1	9.302	9.841	10.948	-1.646	1.107	-0.539
10028005	2	10.444	9.841	10.948	-0.504	1.107	0.603
10028005	3	13.883	9.841	10.948	2.935	1.107	4.042
10028005	4	4.573	9.841	10.948	-6.375	1.107	-5.268
10028005	5	13.769	9.841	10.948	2.820	1.107	3.928
..	..	..	..	..	..	..	..
10028005	13	12.914	9.841	10.948	1.966	1.107	3.073
10060111	1	13.046	9.841	12.953	0.094	3.112	3.205
10060111	2	12.923	9.841	12.953	-0.030	3.112	3.081
10060111	3	13.453	9.841	12.953	0.500	3.112	3.612
10060111	4	13.505	9.841	12.953	0.553	3.112	3.664
10060111	5	12.418	9.841	12.953	-0.535	3.112	2.577

Between- and within-group variation: `xtsum`

- Stata contains a 'canned' routine, xtsum, that summarises within and between variation.
- Doesn't give an exact decomposition:
  - Converts sums of squares to variance using different 'degrees of freedom' so they are not comparable
  - Reports square root (i.e. standard deviation) of these variances
  - Documentation is not very clear!

```
. xtsum wage

Variable      |    Mean   Std. Dev.    Min       Max    |  Obs
--------------+-----------------------------------------+----------
wage  overall | 9.841044  5.712089   .3813552  121.7474 | N = 14547
      between |           4.969431   3.322259  46.54612 | n =  1119
      within  |           2.820121 -18.37394  108.5192 | T =    13
```

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)

```
. tab d1evec if wave==1

 have you ever taken |
             cannabis |      Freq.     Percent        Cum.
----------------------+-----------------------------------
                  yes |        855       25.45       25.45
                   no |      2,477       73.72       99.17
           don't know |         13        0.39       99.55
don't want to answer |         15        0.45      100.00
----------------------+-----------------------------------
                Total |      3,360      100.00
```

Transition matrix

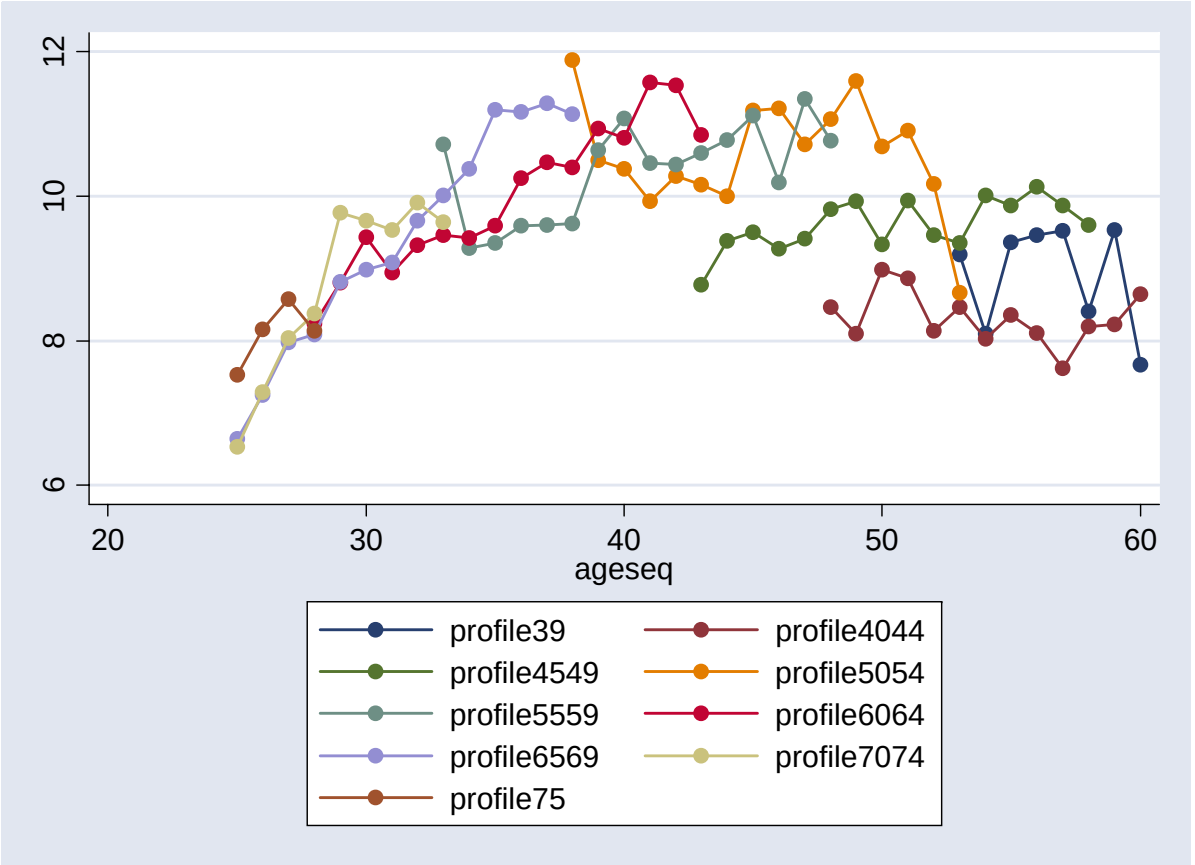
```
. xttrans d1evec, freq
```

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

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

Age and cohort: earnings profiles

How have different generations fared in the labour market?



Day 2:

Approaches to modelling

Basic notation

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

y_{it} = dependent variable to be analysed

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

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

where i indexes individuals, t indexes time periods.

Modelling approaches

Ways of thinking about panel data:

- A collection of cross-sections, one for each time period:
  - Between-group regression
  - The Structural Equations (SEM) approach – 1 equation for each time period (*e.g.* Bollen, 1989, *Structural Equations with Latent Variables*)
- A collection of time-series, one for each individual. Examples:
  - Within-group regression
  - Dynamic models with individual heterogeneity
  - Latent growth curve analysis (*e.g.* Acock & Li http://oregonstate.edu/dept/hdfs/papers/lgcgeneral.pdf#search=%22latent%20growth%20curve%20analysis%20oregon%22)
  - Trajectory analysis (*e.g.* Nagin & Tremblay, *Child Development* 1999)
- Comprehensive models try to capture both inter-individual and inter-period variation

Why use panel data?
The disadvantages of cross-section data

Example: cross-section earnings regression (single time period, t subscript suppressed)

$$y_i = \mathbf{z}_i \alpha + \mathbf{x}_i \beta + \varepsilon_i$$

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")

Some basic identification problems

(1) Unobservable variables

- Can we identify the impact of unobservables?
- 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" $\Rightarrow$ cohort or year-of-birth effect
- Time may affect behaviour through changing social environment

Identification of unobservables

Example: wage models based on human capital theory:

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

where $i = 1 \ldots n$, $t = 1 \ldots T_i$:

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:

Ability (u) is likely to be positively related to education ($\mathbf{z}$)
$\Rightarrow$ upward bias in estimate of returns to education

But can we identify the effect of u_i if we can't observe it?

Identification of unobservables

The identification of the effect of rests on assumptions about the correlation structure of the compound residual v_{it} :

$$v_{it} = u_i + \varepsilon_{it}$$

If individuals have been sampled at random, there is no correlation across different individuals:

$$\text{cov}(u_i , u_j) = 0$$

$$\text{cov}([\varepsilon_{i1} \ldots \varepsilon_{iT}], [\varepsilon_{j1} \ldots \varepsilon_{jT}]) = 0$$

for any two (different) sampled individuals i and j

But there may be some correlation over time for any individual:

$$\text{cov}(v_{is} , v_{it}) \neq 0 \quad \text{for two different periods } s \neq t,$$

since:

$$\text{cov}(v_{is} , v_{it}) = \text{cov}(u_i + \varepsilon_{is} , u_i + \varepsilon_{it}) = \text{var}(u_i) + \text{cov}(\varepsilon_{is} , \varepsilon_{it})$$

If we assume $\text{cov}(\varepsilon_{is} , \varepsilon_{it}) = 0$ then u_i is the only source of correlation over time, so its variance can be identified from the correlation of the residuals.

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

Consider again the panel regression model:

$$y_{it} = \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + u_i + \varepsilon_{it} \qquad (1)$$

Let $\mathbf{z}_i \gamma$ be any arbitrary combination of the z-variables (choose any value for γ you like). Add it to the right-hand side and subtract it again:

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

Now re-write this as:

$$y_{it} = \mathbf{z}_i \alpha^* + \mathbf{x}_{it} \beta + u_i^* + \varepsilon_{it} \qquad (2)$$

Where α^* represents $(\alpha + \gamma)$ and u_i^* represents $(u_i - \mathbf{z}_i \gamma)$.

But (1) and (2) have exactly the same form, so we can't tell whether we're estimating α or a completely arbitrary value $\alpha^* = (\alpha + \gamma)$.

So the separate effects of $\mathbf{z}_i \alpha$ and u_i can't be distinguished empirically without further assumptions

Summary

In models like:

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

- We can only identify the effect of unobservable ability u_i if we can assume that ε_{it} is serially-independent (or some other simple autocorrelation structure).
- We cannot distinguish the separate effects of $\mathbf{z}_i$ and u_i without making further assumptions (*e.g.* no correlation between $\mathbf{z}_i$ and u_i).

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, t varies but A_{it} , t and B_i are collinear - only two of the three effects can be estimated.
- 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 the social environment at the time of birth that generate differences between cohorts and the present social environment that generates time effects.

Let $\mathbf{w}(t)$ be variables describing the social environment at historical time t.

Then our model would use A_{it} , $\mathbf{w}(t)$ and $\mathbf{w}(B_i)$) as covariates

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

When to use regression methods

Regression models are suitable for the analysis of dependent variables y_{it} which can vary continuously, so:

- Income, birthweight, etc. $\Rightarrow$ regression appropriate
- Age at retirement, interpolated grouped income, etc. $\Rightarrow$ regression may work OK
- Age of school leaving, no. of visits to doctor last week, etc. $\Rightarrow$ regression a bit risky
- Binary variables (married/non-married, employed/non-employed, etc. $\Rightarrow$ regression very unreliable

Regression models also have technical problems when:

- The sample is censored or truncated (e.g. if y_{it} = hours of work and non-workers are recorded as zero or excluded)
- When there is no natural scale (e.g. Likert scales)

Related methods (1)

Latent growth curve analysis is widely used in sociology, psychology, criminology, etc. but not economics

Example: simple quadratic latent growth curve:

$$y_{it} = u_i + \alpha_i t + \beta_i t^2 + \varepsilon_{it}$$

where the intercept and slope coefficients (u_i, α_i, β_i) vary randomly across individuals

Advantage:

- Doesn't assume all individuals have the same coefficients (panel data regression assumes no variation in α_i, β_i)

Disadvantage:

- Purely descriptive: no theory of development
- Crude dynamics (nothing changes the trend for an individual once it's underway)

Related methods (2)

Structural equation modelling (SEM) is widely used in psychology and economics, but with differences in terminology.

In panel data applications, each year is described by a different equation:

Period 1: $y_{i1} = \mathbf{z}_i \alpha_1 + \mathbf{x}_{i1} \beta_1 + u_i + \varepsilon_{i1}$

.
.

Period T: $y_{iT} = \mathbf{z}_i \alpha_T + \mathbf{x}_{iT} \beta_T + u_i + \varepsilon_{iT}$

Advantage:

- general structure (e.g. panel regression is special case where the α_t and β_t are the same in all periods)

Disadvatage:

- No theory of how the parameters vary over time
- Can't predict outcomes in new periods
- Difficult to use in long or very unbalanced panels

Related methods (3)

Multi-level modelling is widely used throughout social statistics. It generalises ordinary panel data applications to multiple dimensions

Example: time periods *(t)* within individuals (*i*) within households (*h*):

$$y_{hit} = \mathbf{x}_{hit}\beta + u_{hi} + w_h + \varepsilon_{iT}$$

- w_h is the household effect, common to all individuals at all periods within household *h*
- u_{hi} is the individual effect, common to all time periods for the *i*th individual in household *h*

Specialist software is available for latent growth curve, SEM and Multi-level analysis (MLwin, Mplus, LISREL, etc). See also *xtmixed* and GLLAMM in Stata

Pooled regression for panel data

The "standard" panel data regression model is:

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

We have observations indexed by $t = 1 \ldots T_i$, $i = 1 \ldots n$.

- A pooled regression of y on $\mathbf{z}$ and $\mathbf{x}$ using all the data together would assume that there is no correlation across individuals, nor across time periods for any individual
- This would ignore the individual effect u, which generates correlation between the values of $(u_i + \varepsilon_{i1}) \ldots (u_i + \varepsilon_{iT})$ for each individual i
- So pooled regression doesn't make best use of the data
  - Under favourable conditions (if u_i is uncorrelated with $\mathbf{z}_i$ and $\mathbf{x}_{it}$), pooled regression gives unbiased but inefficient results, with incorrect standard errors, t-ratios, etc.
  - If u_i is correlated with $\mathbf{z}_i$ and $\mathbf{x}_{it}$, pooled regression is also biased

Least-squares dummy variable (LSDV) regression

The panel data regression model is:

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

We have observations indexed by $t = 1 \ldots T_i$, $i = 1 \ldots n$.

The u_i can be captured using dummy variables. Construct a set of n dummy variables $D1_i \ldots Dn_i$, where:

$$Dr_i = 1 \text{ if } i = r \text{ and 0 otherwise, for } r = 1 \ldots n$$

Thus Dr_{it} tells us whether observation i, t relates to person r.

The model is now:

$$y_{it} = \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_1 D1_i + \ldots + u_n Dn_i + \varepsilon_{it}$$

So $u_1 \ldots u_n$ are now seen as the coefficients of a set of n dummy variables.

Shortcut calculation of the LSDVregression

A multiple regression of y on ($\mathbf{z}$, $\mathbf{x}$) and ($D1$ … Dn) can be done in two stages:

Stage 1: Eliminate the effect of ($D1$ … Dn) on each of the variables (y, $\mathbf{z}$, $\mathbf{x}$) using the "within-group" data transformation:

$$y_{it}^{*} = y_{it} - \bar{y}_i$$

$$\mathbf{x}_{it}^{*} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$$

$$\mathbf{z}_i^{*} = \mathbf{z}_i - \bar{\mathbf{z}}_i \equiv \mathbf{0} \qquad \text{(so } \mathbf{z}_i \text{ is eliminated completely)}$$

Stage 2: regress y^{*} on ($\mathbf{z}^{*}$, $\mathbf{x}^{*}$) : in other words, $y_{it} - \bar{y}_i$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$

[Intuition: think of regressing a variable on a constant. Estimate of constant is mean and residual is deviation from mean.]

This is exactly equivalent to regressing y on ($\mathbf{z}$, $\mathbf{x}$) and ($D1$ … Dn)

Another interpretation of LSDV

Start differently, by thinking how we can cope with u_i

We don't know its statistical properties, so let's try to eliminate it from the model. We can eliminate it in various ways, for example:

Time differencing: $y_{it} - y_{it-1} = (\mathbf{x}_{it} - \mathbf{x}_{it-1})\boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{it-1}$

or

Within-group transform: $y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i$

The within-group approach is the most efficient in the least squares sense.

A note on terminology

Different names are commonly used for this one estimation method:

- Least squares dummy variables (LSDV)
- Within-group regression
- Fixed-effects regression
- Covariance analysis regression

$\Rightarrow$ "LSDV" refers to the method of derivation using explicit dummy variables;

$\Rightarrow$ "within-group" refers to the type of data transform implied by the method;

$\Rightarrow$ "fixed effects" is common but often poor terminology which suggests (wrongly, in the case of sample survey data) that the u_i are fixed parameters

$\Rightarrow$ "covariance analysis" reflects the origins of the method as a generalisation of analysis of variance in agricultural experiments

Between-group regression

Instead of eliminating u_i from the regression, we can amplify it by averaging out all the within-individual variation, leaving only between-individual variation to analyse:

Between-group transform: $\bar{y}_i = \mathbf{z}_i\boldsymbol{\alpha} + \bar{\mathbf{x}}_i\boldsymbol{\beta} + u_i + \bar{\varepsilon}_i$

Then regress $\bar{y}_i$ on $(\mathbf{z}_i, \bar{\mathbf{x}}_i)$ in one of two ways:

- Use one group-mean observation per individual
- Use T_i copies of the group mean data for individual i

Note: The latter is equivalent to a weighted regression of $\bar{y}_i$ on $\bar{\mathbf{x}}_i$, with a weight of T_i for individual i. It is often desirable to give more weight to individuals with many time observations.

Within- & between-group estimates – simple case

Suppose that x (and therefore β) is a single variable (scalar), and panel is balanced ($T_i = T$). Want to estimate:

Within-group: $y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + \varepsilon_{it} - \bar{\varepsilon}_i$

Between-group: $\bar{y}_i = \bar{x}_i\beta + u_i + \bar{\varepsilon}_i$

$$\hat{\beta}_W = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)^2} \equiv \frac{w_{xy}}{w_{xx}} \quad ; \quad \hat{\beta}_B = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{\sum_{i=1}^{n}\sum_{t=1}^{T}(\bar{x}_i - \bar{x})^2} \equiv \frac{b_{xy}}{b_{xx}}$$

Within-group estimate – simple case

Can substitute for $y_{it} - \bar{y}_i$ in preceding formula, to obtain:

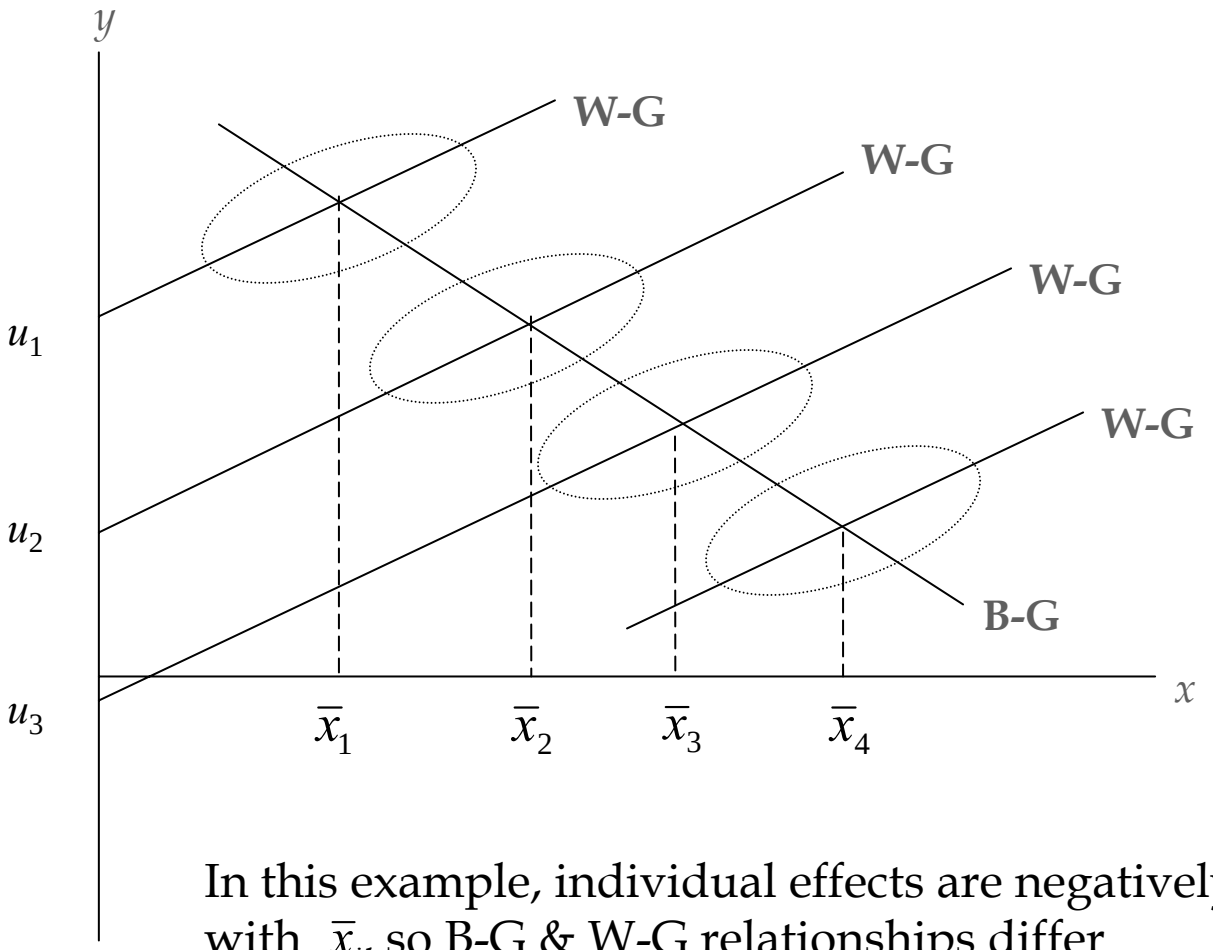
$$\hat{\beta}_W = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(\{(x_{it} - \bar{x}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)\})}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)^2} = \beta + \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i)}{\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)^2} \equiv \beta + \frac{w_{x\varepsilon}}{w_{xx}}$$

If x_{it} and ε_{it} are uncorrelated, $E(w_{x\varepsilon}) = 0$, so $E\hat{\beta}_W = \beta$

…which means, loosely speaking, that on average $\hat{\beta}_W$ is correct (unbiased).

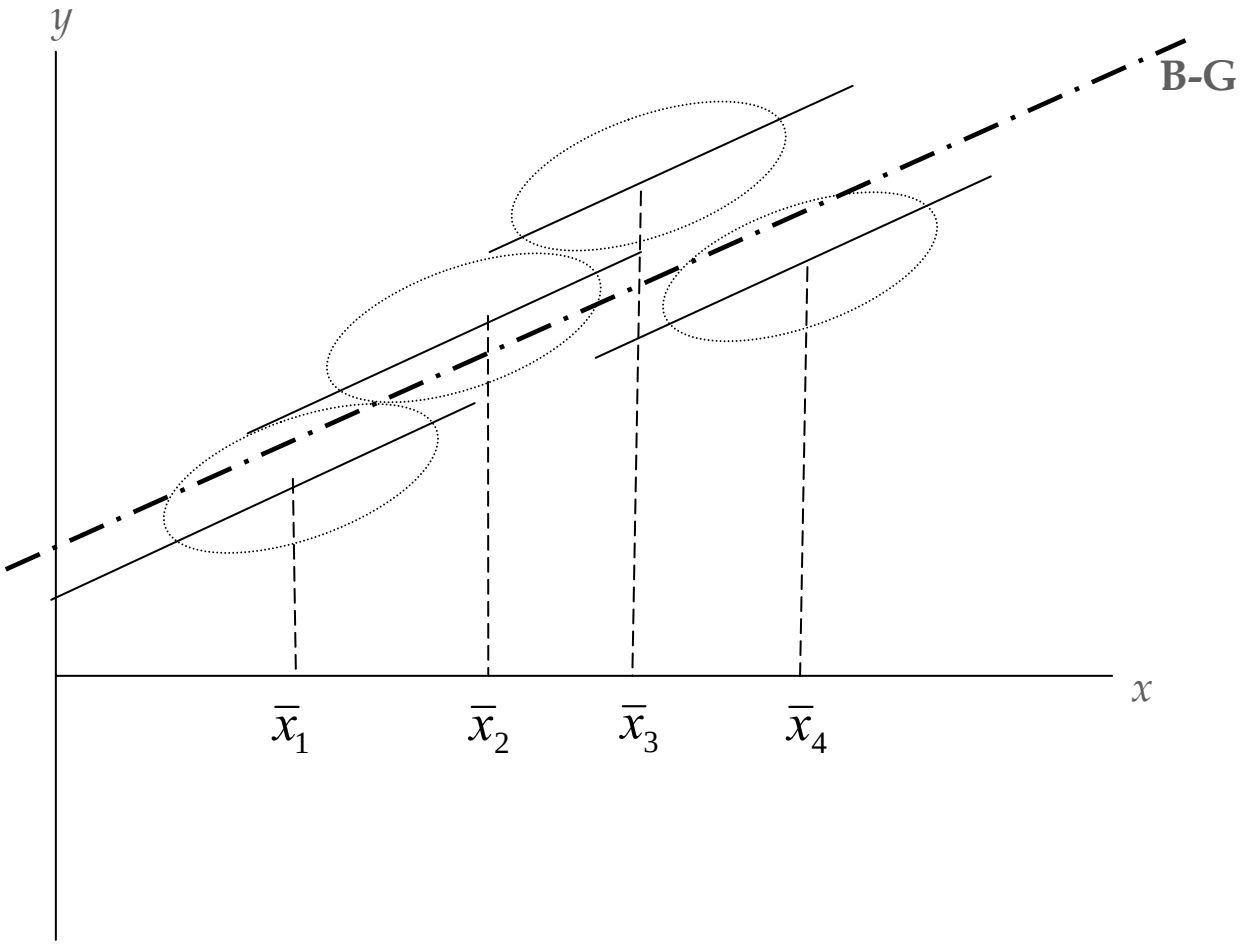
Note: for unbiasedness of $\hat{\beta}_B$, we need also that x_{it} is uncorrelated with u_i $\Rightarrow$ so within-group regression is less "robust"

Within- & between-group relationships:
correlated individual effects



In this example, individual effects are negatively correlated with $\bar{x}_i$, so B-G & W-G relationships differ

Within- & between-group relationships:
uncorrelated individual effects



Example of panel data estimation

The Stata command *xtreg* computes within-group and between-group regressions

Example: within- and between-group regressions of log earnings on age, year of birth and time, allowing for unobserved individual effects:

gen age=year-cohort

gen lwage=ln(w_hr)

xtreg lwage age cohort, fe

xtreg lwage age cohort, be

Stata output: within-group regression

```
. xtreg lwage age cohort , fe

Fixed-effects (within) regression               Number of obs      =     61516
Group variable (i): pid                         Number of groups   =     10335

R-sq:  within  = 0.1217                         Obs per group: min =         1
       between = 0.0312                                        avg =       6.0
       overall = 0.0194                                        max =        14

                                                F(1,51180)         =   7094.59
corr(u_i, Xb)  = -0.4880                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .030061   .0003569    84.23   0.000     .0293615    .0307605
      cohort |  (dropped)
       _cons |   .8994719     .01369    65.70   0.000     .8726394    .9263045
-------------+----------------------------------------------------------------
     sigma_u |  .60455798
     sigma_e |  .28494801
         rho |  .81822708   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(10334, 51180) =    18.19        Prob > F = 0.0000
```

Stata output: between-group regression

```
. xtreg lwage age cohort , be

Between regression (regression on group means)  Number of obs      =     61516
Group variable (i): pid                         Number of groups   =     10335

R-sq:  within  = 0.1217                         Obs per group: min =         1
       between = 0.0356                                        avg =       6.0
       overall = 0.0313                                        max =        14

                                                F(2,10332)         =    190.55
sd(u_i + avg(e_i.))=  .5277749                  Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0188575   .0017201    10.96   0.000     .0154858    .0222292
      cohort |   .0105401   .0015325     6.88   0.000     .0075361    .0135442
       _cons |  -19.39964   3.065617    -6.33   0.000    -25.40885   -13.39044
------------------------------------------------------------------------------
```

Important points

- The within-group R^2 is much higher than the between-group R^2

  $\Rightarrow$ the covariate *age* "explains" a reasonable amount of the pay variation over time for a given individual

  $\Rightarrow$ but pay differences between individuals are less closely related to age and cohort in R^2 terms

- The large coefficient differences between the within- and between-group age coefficients suggest that a single regression model with classical assumptions doesn't fit the evidence very well

Technical appendix

The following slides can be safely ignored if you're not interested in technical detail or if you aren't familiar with vector-matrix notation and matrix algebra

Coefficient estimates – general formula

In matrix form, the within-group multiple regression is:

$$\hat{\boldsymbol{\beta}} = \mathbf{W}_{xx}^{-1}\mathbf{w}_{xy} = \boldsymbol{\beta} + \mathbf{W}_{xx}^{-1}\mathbf{w}_{x\varepsilon}$$

where $\mathbf{W}_{xx}$, $\mathbf{w}_{xy}$ and $\mathbf{w}_{x\varepsilon}$ are within-group moment matrices:

$$\mathbf{W}_{xx} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)'\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)$$

$$\mathbf{w}_{x\varepsilon} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)'\left(\varepsilon_{it} - \bar{\varepsilon}_i\right)$$

If $\mathbf{x}_{it}$ and ε_{it} are uncorrelated, $E(\mathbf{w}_{x\varepsilon}) = \mathbf{0}$, so:

$$E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$$

Residuals

There are two residuals for the within-group regression:

$$\hat{e}_i = \bar{y}_i - \bar{\mathbf{x}}_i\hat{\boldsymbol{\beta}}$$

$$\hat{\varepsilon}_{it} = \left(y_{it} - \bar{y}_i\right) - \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\beta}} = y_{it} - \mathbf{x}_{it}\hat{\boldsymbol{\beta}} - \hat{e}_i$$

$\hat{e}_i$ is an estimate of $\mathbf{z}_i\boldsymbol{\alpha} + u_i$; $\hat{\varepsilon}_{it}$ is an estimate of ε_{it}

Since $\hat{\varepsilon}_{it}$ is the residual from the LSDV multiple regression, its variance is an unbiased estimator of σ_ε^2 under the classical assumptions of independent sampling of individuals and:

$$E\varepsilon_{it} = 0; \quad E\varepsilon_{it}^{\ 2} = \sigma_\varepsilon^2$$

$$E\mathbf{x}_{is}\varepsilon_{it} = \mathbf{0} \quad \text{for all } i, s, t$$

$$E\varepsilon_{is}\varepsilon_{it} = 0 \quad \text{for all } i, s \neq t$$

Estimation of α

The residual $\hat{e}_i$ can be written:

$$\hat{e}_i = \bar{y}_i - \bar{\mathbf{x}}_i\hat{\boldsymbol{\beta}} = \left(\mathbf{z}_i\boldsymbol{\alpha} + \bar{\mathbf{x}}_i\boldsymbol{\beta} + u_i + \bar{\varepsilon}_i\right) - \bar{\mathbf{x}}_i\hat{\boldsymbol{\beta}}$$
$$= \mathbf{z}_i\boldsymbol{\alpha} + u_i + \bar{\varepsilon}_i - \bar{\mathbf{x}}_i\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

Since $\hat{e}_i$ is an estimate of $\mathbf{z}_i\boldsymbol{\alpha} + u_i$, we could regress it on $\mathbf{z}_i$ to estimate $\boldsymbol{\alpha}$. (Use T_i repeated observations on the group means for individual i, to weight individuals appropriately). This gives:

$$\hat{\boldsymbol{\alpha}} = \mathbf{B}_{zz}^{-1}\mathbf{b}_{z\hat{e}}$$

where $\mathbf{B}_{xx}$ *etc.* are between-group cross-product matrices:

$$\mathbf{B}_{zz} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{z}_i'\mathbf{z}_i \; ; \quad \mathbf{b}_{z\hat{e}} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{z}_i'\hat{e}_i$$

Estimation of $\hat{\boldsymbol{\alpha}}$

Rewrite $\hat{\boldsymbol{\alpha}}$ as:

$$\hat{\boldsymbol{\alpha}} = \mathbf{B}_{zz}^{-1}\mathbf{b}_{z\hat{e}} = \boldsymbol{\alpha} + \mathbf{B}_{zz}^{-1}\mathbf{b}_{zu} + \mathbf{B}_{zz}^{-1}\mathbf{b}_{z\varepsilon} - \mathbf{B}_{zz}^{-1}\mathbf{B}_{zx}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

But $\hat{\boldsymbol{\beta}}$ is unbiased and we assume $\mathbf{z}_i$ is uncorrelated with ε_{it} , so:

$$E\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + E\left(\mathbf{B}_{zz}^{-1}\mathbf{b}_{zu}\right)$$

Thus $\hat{\boldsymbol{\alpha}}$ is only unbiased if u_i and $\mathbf{z}_i$ are uncorrelated.

Estimation of σ_u^2

One way is to use the between-group regression. Replace each observation by the individual mean:

$$\bar{y}_i = \mathbf{z}_i\boldsymbol{\alpha} + \bar{\mathbf{x}}_i\boldsymbol{\beta} + u_i + \bar{\varepsilon}_i\,, \qquad i = 1...n;\, t = 1...T_i$$

Estimator: $$\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{zz} & \mathbf{B}_{zx} \\ \mathbf{B}_{xz} & \mathbf{B}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{b}_{zy} \\ \mathbf{b}_{xy} \end{pmatrix}$$

The residual variance, s_B^2, is an estimate of $\sigma_u^2 + \sigma_\varepsilon^2 / \bar{T}$ so:

$$\hat{\sigma}_u^2 = s_B^2 - \frac{s_W^2}{\bar{T}}$$

where s_B^2 and s_W^2 are the b-g and w-g residual variances and $\bar{T}$ is the mean no. of observations per individual.

Note that $\hat{\sigma}_u^2$ may be negative! (If so, Stata sets it to zero!)

Asymptotics for short panels

For panel data arising from repeated surveys, n is usually much larger than $T = \max(T_i)$. This suggests using asymptotic theory based on $n \to \infty$, with all T_i fixed.

Incidental parameters problem: If we regard the unobserved effects $u_1 \ldots u_n$ as parameters to be estimated, then the dimension of the parameter space $\to \infty$ as $n \to \infty$. Standard asymptotic theory doesn't work in this case.

Consistency of within-group estimator:

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}}_W = \boldsymbol{\beta} + \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right)^{-1}$$

$$\times \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\varepsilon_{it} - \bar{\varepsilon}_i) \right)$$

$$= \boldsymbol{\beta} + \left(\operatorname*{plim}_{n\to\infty} \mathbf{W}_{xx} \right)^{-1} \times \mathbf{0} \quad = \boldsymbol{\beta}$$

Day 3: Linear regression analysis: random effects

- Random effects regression: Testing the FE and RE assumptions
  - The Hausman test
  - The Mundlak approach
- Endogeneity issues
  - Forms of endogeneity
  - Endogenous regressors: the between and within-group IV estimator
  - Correlated individual effects: Hausman-Taylor estimation

'Random effects' GLS & ML estimation

- In general, since individuals are sampled at random from the population, u_i (and all other variables) are random: so "random effects" is tautological
- Extract the overall mean from u_i:

$$y_{it} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

- Use $\mathbf{X}_i$ as shorthand for the person i's time series $\{ \mathbf{x}_{i1} \ldots \mathbf{x}_{iT} \}$
- We may choose to assume that u_i is uncorrelated with $\mathbf{z}_i$ and $\mathbf{X}_i$:

$$\mathrm{E}(u_i \mid \mathbf{z}_i, \mathbf{X}_i) = 0 \quad \Rightarrow \mathrm{cov}(u_i, \mathbf{z}_i) = 0 \;\&\; \mathrm{cov}(u_i, \mathbf{X}_i) = 0$$

- Assume also homoskedasticity and uncorrelatedness

$$\mathrm{E}(u_i^2 \mid \mathbf{z}_i, \mathbf{X}_i) = \sigma_u^2 \;;\; \mathrm{E}(u_i \varepsilon_{it} \mid \mathbf{z}_i, \mathbf{X}_i) = 0 \quad \text{for all } t$$

- Then write the composite random disturbance as:

$$v_{it} = u_i + \varepsilon_{it}$$

- What is the covariance structure of the random process $\{ v_{it} \}$?

Random effects covariance structure

Variances & covariances (conditional on $\mathbf{z}_i$, $\mathbf{X}_i$) :

$$\text{var}(v_{it}) = \sigma_u^2 + \sigma_\varepsilon^2 ; \quad \text{cov}(v_{it} , v_{is}) = \sigma_u^2 \quad \text{for all } s \neq t$$

So the observations from different time periods (and the same individual) are not independent: they are *equi-correlated*.

The observations are *clustered* by individual, with non-zero *intra-group* correlations

The positive correlation between observations for any individual means that within-person variation is less than it would otherwise be. Consequently, whatever within-person variation we do have is particularly informative

$\Rightarrow$ give more weight to within- than between-group variation

Generalised Least Squares

Generalised least squares (GLS) does this weighting for us.

For simplicity, assume just one explanatory variable, $\mathbf{x}_{it}$. Then GLS is:

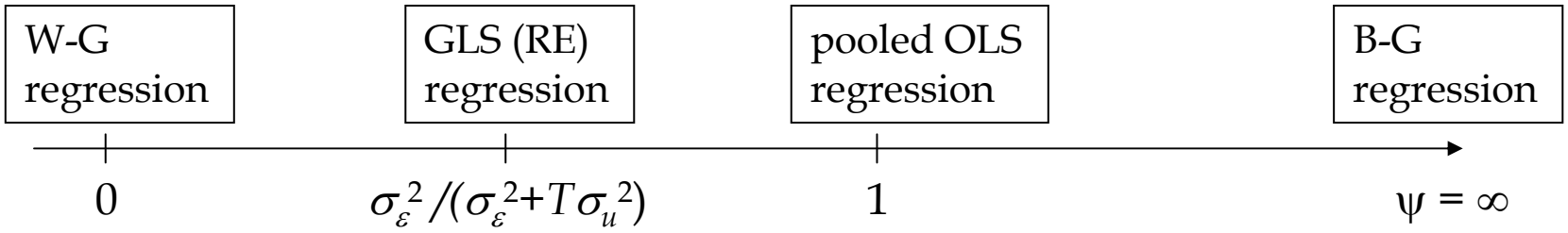
$$\hat{\beta}_{GLS} = \frac{\sum_{i=1}^{n}\left[w_{xyi} + \psi_i b_{xyi}\right]}{\sum_{i=1}^{n}\left[w_{xxi} + \psi_i b_{xxi}\right]}$$

where:

$$\psi_i = \sigma_\varepsilon^2 / \left(\sigma_\varepsilon^2 + T_i \sigma_u^2\right)$$

$$w_{xxi} = \sum_{t=1}^{T_i}\left(x_{it} - \bar{x}_i\right)^2, \quad b_{xxi} = T_i\left(\bar{x}_i - \bar{x}\right)^2 \qquad \text{etc.}$$

Estimators combining within & between-group variation



- If σ_ε^2 is zero, then GLS is the same as w-g regression
- If σ_u^2 is zero, then GLS is the same as pooled OLS
- GLS is never the same as b-g regression (since $\sigma_\varepsilon^2/(\sigma_\varepsilon^2+T\sigma_u^2)$ can't be greater than 1) $\Rightarrow$ b-g regression is never an efficient method

GLS properties

Note that :

- GLS uses the optimal (efficient) combination of within and between variation: OLS (i.e. with $\psi_i = 1$) is *not* generally the efficient estimator.
- $\psi_i < 1$, so less weight is given to between-group variation
- $\lim_{T_i \to \infty} \psi_i = 0$, so between-group variation is unimportant in a long panel, and the GLS estimator converges to the within estimator, i.e. $\hat{\boldsymbol{\beta}}_{GLS} \to \hat{\boldsymbol{\beta}}_{W}$ as the panel lengthens
- If individual effects do not matter ($\sigma_u^2 = 0$) then $\psi_i = 1$ and it is easily shown that $\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_{OLS}$

Feasible GLS

We can only use GLS if we know the variance parameters σ_ε^2 and σ_u^2 . They can be estimated from the within-group and between-group regression residuals.

Consider the full regression model involving both **z** and **x**. It can be shown that GLS is equivalent to the following procedure:

(1) Transform the data:

$$y_{it}^{+} = y_{it} - \theta_i \bar{y}_i \; ; \quad \mathbf{z}_i^{+} = (1-\theta_i)\mathbf{z}_i \; ; \quad \mathbf{x}_{it}^{+} = \mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i$$

where:

$$\theta_i = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T_i \sigma_u^2}}$$

(2) Regress y_{it}^{+} on $\left(\mathbf{z}_i^{+}, \mathbf{x}_{it}^{+}\right)$, pooling all observations

Maximum likelihood

Speaking loosely, the likelihood function measures the degree to which our model is consistent with the data, at any particular choice of values for the model parameters. So we can estimate all the parameters (α , β , σ_ε^2 , σ_u^2) together by choosing their values to maximise the likelihood function (see appendix for details).

ML and feasible GLS are statistically equivalent if n is very large.

In Stata, the command *xtreg* has various options:

,fe for within-group
,be for between-group
,re for random effects (feasible GLS)
,mle for random effects (ML)

Fixed effects or random effects? Concepts and interpretation

- Specification of model as FE or RE depends partly on the nature of data. For example:
  - If individuals are randomly sampled from population then u_i is random (a 'draw' from the population distribution).
  - But for an industry level analysis, where we observe a panel of all industries over several years, industry effect u_i can be thought of as a fixed effect.
- RE implies *unconditional* inference (because we don't want to be restricted to the particular individuals sampled), while FE implies inference *conditional* on the effects in the sample.
- In practice, with randomly sampled data, FE/RE choice is based on whether a further assumption holds: that u_i is uncorrelated with the regressors: $E(u_i \mid \mathbf{z}_i, \mathbf{X}_i) = 0$

Testing the hypothesis of uncorrelated effects

The random effects estimator (and any estimator that uses between-group variation) is only unbiased (strictly, consistent as $n \to \infty$) if the following hypothesis is true:

$$H_0: \quad \mathrm{E}(u_i \mid \mathbf{z}_i, \mathbf{X}_i) = 0$$

It is important to test H_0. There are various equivalent ways of doing so, including:

(1)Hausman test: is the difference $\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{GLS}$ large?

(2)Between-within comparison: is $\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_B$ large?

(3) Mundlak approach: estimate the model

$$y_{it} = \alpha_0 + \mathbf{z}_i\boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + \eta_i + \varepsilon_{it}$$

by GLS and test H_0: $\boldsymbol{\gamma} = \mathbf{0}$.

Hausman test

The idea of the Hausman test is to compare two estimators which should be approximately the "same" if the zero-correlation assumption holds (H_0), but different if the assumption is false (H_1).

Specifically, under H_0 both estimators $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_{GLS}$ are unbiased (strictly, consistent), and $\hat{\boldsymbol{\beta}}_{GLS}$ is more efficient, (so $\text{var}(\hat{\boldsymbol{\beta}}_W) > \text{var}(\hat{\boldsymbol{\beta}}_{GLS})$).

It can be shown that the variance (matrix) of $\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{GLS}$ is:

$$\text{var}\left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{GLS}\right) = \text{var}\left(\hat{\boldsymbol{\beta}}_W\right) - \text{var}\left(\hat{\boldsymbol{\beta}}_{GLS}\right)$$

Under H_1, $\hat{\boldsymbol{\beta}}_W$ is still unbiased but $\hat{\boldsymbol{\beta}}_{GLS}$ is not. So the Hausman test statistic:

$$S = \left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{GLS}\right)' \left[\text{var}\left(\hat{\boldsymbol{\beta}}_W\right) - \text{var}\left(\hat{\boldsymbol{\beta}}_{GLS}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_{GLS}\right)$$

should take a large value and reject if H_0 is not true.

If H_0 is true, the statistic S is approximately distributed as χ^2 with k d.f. where k = number of variables in $\mathbf{x}_{it}$, so we use critical values for the $\chi^2(k)$ distribution.

BHPS example: feasible GLS estimates

```
. xtreg lwage age cohort , re

Random-effects GLS regression                   Number of obs      =     61516
Group variable (i): pid                         Number of groups   =     10335

R-sq:  within  = 0.1217                         Obs per group: min =         1
       between = 0.0335                                        avg =       6.0
       overall = 0.0345                                        max =        14

Random effects u_i ~ Gaussian                   Wald chi2(2)       =   7405.63
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0295982   .0003497    84.63   0.000     .0289127    .0302836
      cohort |   .0201181   .0004723    42.60   0.000     .0191924    .0210438
       _cons |  -38.56221   .9343531   -41.27   0.000    -40.39351   -36.73091
-------------+----------------------------------------------------------------
     sigma_u |  .49751772
     sigma_e |  .28495079
         rho |  .75299116   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

BHPS example: within-group estimates

```
. xtreg lwage age cohort , fe

Fixed-effects (within) regression               Number of obs      =     61516
Group variable (i): pid                         Number of groups   =     10335

R-sq:  within  = 0.1217                         Obs per group: min =         1
       between = 0.0312                                        avg =       6.0
       overall = 0.0194                                        max =        14

                                                F(1,51180)         =   7094.59
corr(u_i, Xb)  = -0.4880                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .030061   .0003569    84.23   0.000     .0293615    .0307605
      cohort |  (dropped)
       _cons |   .8994719     .01369    65.70   0.000     .8726394    .9263045
-------------+----------------------------------------------------------------
     sigma_u |  .60455798
     sigma_e |  .28494801
         rho |  .81822708   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(10334, 51180) =    18.19        Prob > F = 0.0000
```

Example: BHPS Hausman test

```
. hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+------------------------------------------------------------
         age |     .030061     .0295982        .0004628        .0000711
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       42.34
                Prob>chi2 =      0.0000
```

Conclusion: we reject H_0 – there is correlation between u_i and age, so the within-group regression is biased

But note: although the FE-RE difference is statistically significant, it is rather small

The Mundlak approach

Mundlak (1978) suggested that a way to reconcile FE and RE models was to approximate the individual effect as a function of the individual means of time-varying characteristics:

$$u_i = \bar{\mathbf{x}}_i \boldsymbol{\gamma} + \eta_i$$

Substituting into the main model:

$$y_{it} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\gamma} + \eta_i + \varepsilon_{it}$$

Estimating by GLS yields $\hat{\boldsymbol{\beta}}_{M,GLS} \equiv \hat{\boldsymbol{\beta}}_W$ because the (linear) dependence of u_i on $\mathbf{x}_{it}$ is fully captured by the Mundlak formulation [note this is not true for non-linear models, as we see later].

A test of $\text{cov}(u_i , \mathbf{x}_{it}) = 0$ is a test of H_0: $\boldsymbol{\gamma} = \mathbf{0}$.

- If the test rejects H_0, GLS using the un-augmented RE model (without $\bar{\mathbf{x}}_i$) is biased $\Rightarrow$ we should use the FE model.
- If the test doesn't reject H_0, $\Rightarrow$ we should use GLS on the original model.

Example: Mundlak test

```
. xtreg lwage age cohort mage, re

Random-effects GLS regression                   Number of obs      =     61516
Group variable (i): pid                         Number of groups   =     10335

R-sq:  within  = 0.1217                         Obs per group: min =         1
       between = 0.0356                                        avg =       6.0
       overall = 0.0370                                        max =        14

Random effects u_i ~ Gaussian                   Wald chi2(3)       =   7453.30
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .030061   .0003567    84.27   0.000     .0293618    .0307601
      cohort |   .0103154   .0015734     6.56   0.000     .0072317    .0133992
       mage |  -.0117292   .0017958    -6.53   0.000     -.015249   -.0082095
       _cons |  -18.93191   3.147336    -6.02   0.000    -25.10057   -12.76324
-------------+----------------------------------------------------------------
     sigma_u |  .49751772
     sigma_e |  .28495079
         rho |  .75299116   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Endogeneity

- Forms of endogeneity
- Endogenous regressors: the between and within-group IV estimator
- Correlated individual effects: Hausman-Taylor estimation

Endogeneity in static models

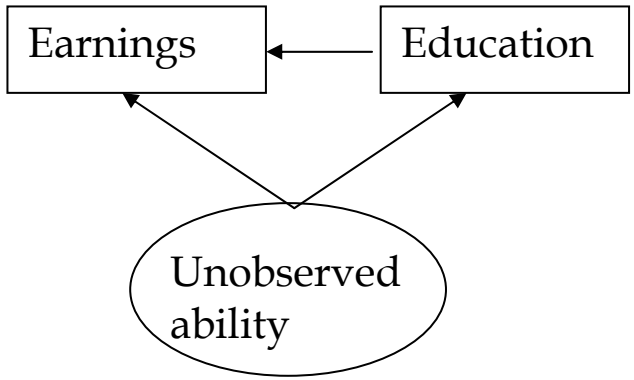
Example: an earnings model

$$y_{it} = \alpha_1 Educ_i + \alpha_2 Female + \beta_1 Age_{it} + \beta_2 Tenure_{it} + u_i + \varepsilon_{it}$$

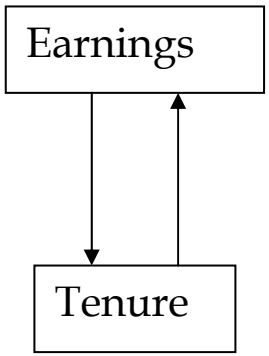
Two forms of endogeneity:

Two-way causation: experience is rewarded with high pay & workers tend to stay in high-paid jobs

Unobserved common factors: ability is rewarded with high pay & high-ability people stay longer in education



(a) unobserved common factor



(b) 2-way causation

Example of endogeneity

Example: an earnings model

$$y_{it} = \alpha_1 Educ_i + \alpha_2 Female + \beta_1 Age_{it} + \beta_2 Tenure_{it} + u_i + \varepsilon_{it}$$

(1) Two-way causation: workers tend to stay in high-paid jobs:

Tenure model: $Tenure_{it} = \gamma y_{it} + \upsilon_{it} \qquad (\gamma > 0)$

$= \gamma(\alpha_1 Educ_i + \ldots + \beta_1 Age_{it} + \beta_2 Tenure_{it} + u_i + \varepsilon_{it}) + \upsilon_{it}$

$= [\gamma(\alpha_1 Educ_i + \ldots + \beta_1 Age_{it} + u_i + \varepsilon_{it}) + \upsilon_{it}] / (1 - \gamma\beta_2)$

$\Rightarrow \quad \text{cov}(Tenure_{it}, u_i) = \gamma\sigma_u^2 / (1 - \gamma\beta_2)$

$\text{cov}(Tenure_{it}, \varepsilon_{it}) = \gamma\sigma_\varepsilon^2 / (1 - \gamma\beta_2)$

(2) Unobserved common factors: u_i represents ability & high-ability people stay longer in education:

$Educ_i = \delta u_i + \text{other vars} \qquad (\delta > 0)$

$\Rightarrow \quad \text{cov}(Educ_i, u_i) = \delta\sigma_u^2$

$\text{cov}(Educ_i, \varepsilon_{it}) = 0$

Strategy for dealing with endogeneity

Type of endogeneity	Consequences	Method
2-way causation (e.g. tenure → wage & wage → tenure)	$\text{Cov}(x,u) \neq 0$ $\text{Cov}(x,\varepsilon) \neq 0$	Within-group IV (w-g to eliminate u_i and IV to deal with covariance with ε)
Common unobserved factor which persists over time (e.g. ability → wage, ability → education & education → wage)	$\text{Cov}(x,u) \neq 0$ $\text{Cov}(x,\varepsilon) = 0$	Within-group regression (eliminates u_i) and Hausman-Taylor to estimate coefficients of $\mathbf{z}_i$
Common unobserved factor which does not persist over time (e.g. job loss → wage & job loss → tenure)	$\text{Cov}(x,u) = 0$ $\text{Cov}(x,\varepsilon) \neq 0$	Random-effects IV, using as IVs variables which are correlated with risk of job loss but not wages; no need to use within-group, since u_i isn't correlated with x
None	$\text{Cov}(x,u) = 0$ $\text{Cov}(x,\varepsilon) = 0$	GLS random effects regression

The Instrumental Variables principle

Simple example – a cross-section regression model:

$$y_i = x_i \beta + \varepsilon_i$$

Problem: simultaneous causation

$\Rightarrow \text{cov}(x_i , \varepsilon_i) \neq 0$

$\Rightarrow$ OLS regression of y_i on x_i is biased

But assume there is another variable q_i with two properties:

Validity: $\text{cov}(q_i , \varepsilon_i) = 0$

Relevance: $\text{cov}(q_i , x_i) \neq 0$

The *validity* requirement says that the instrument must not suffer from the same endogeneity problem that x_i does;

The *relevance* requirement says that the instrument must be closely related to x_i

Motivation for the IV method

The assumption of instrument validity is a ***moment condition*** which states that a particular *moment*, cov(q, ε), must be equal to zero

But the model tells us that: $\varepsilon_i = y_i - x_i \beta$, so:

$$\begin{aligned} \text{cov}(q_i , \varepsilon_i) &= \text{cov}(q_i , [\, y_i - x_i \beta \,]) \\ &= \text{cov}(q_i , y_i) - \beta \,\text{cov}(q_i , x_i) \\ &= 0 \quad \text{(instrument validity requirement)} \end{aligned}$$

Solve for β:

$$\beta = \text{cov}(q_i , y_i) \,/\, \text{cov}(q_i , x_i)$$

So, if q is a valid instrument, β must be equal to the ratio of the population covariance between q and y and between q and x.

The simple Instrumental Variable (IV) estimator

The sample analogue of this moment condition provides an estimator:

$$\hat{\beta}_{IV} = \frac{\text{sample cov}(q, y)}{\text{sample cov}(q, x)} = \frac{\sum_{i=1}^{n}(q_i - \bar{q})(y_i - \bar{y})}{\sum_{i=1}^{n}(q_i - \bar{q})(x_i - \bar{x})}$$

This can be generalised to:

- More than one explanatory variable in ($\mathbf{z}_i$, $\mathbf{x}_{it}$)
- More than one instrumental variable
- But we must have number of instruments ≥ number of explanatory variables

(See technical appendix)

Simultaneity: Within-group IV estimation

Model:

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

Partition $\mathbf{x}_{it}$:

$$\mathbf{x}_{it} = (\mathbf{x}_{1it}, \mathbf{x}_{2it}),$$

Where $\mathbf{x}_{2it}$ represents the endogenous covariates:

$$\text{cov}(\mathbf{x}_{1it}, \varepsilon_{it}) = 0 \text{ and } \text{cov}(\mathbf{x}_{2it}, \varepsilon_{it}) \neq 0$$

Find a set of instruments $\mathbf{q}_{2it}$ (at least as many as in $\mathbf{x}_{2it}$)

where $\text{cov}(\mathbf{q}_{2it}, \varepsilon_{it}) = 0$

Full set of instruments: $\mathbf{q}_{it} = (\mathbf{x}_{1it}, \mathbf{q}_{2it})$

Within-group transformation:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i$$

Within-group IV estimator uses $(\mathbf{q}_{it} - \bar{\mathbf{q}}_i)$ as instruments

Other IV estimators

- By applying the between-group transform or the random-effects GLS transform to the model and instruments, we can define between-group and random effects IV estimators analogous to the regression case.
- Like regression, these are not robust with respect to correlation between u_i and $(\mathbf{z}_i , \mathbf{x}_{it})$
- So the Random Effects IV method should only be used if we think the endogeneity problem arises because of the presence of non-persistent common unobserved factors (i.e. ε_{it}) influencing both y and x. If there are also common persistent factors (i.e. u_i), then RE-IV will be biased
- See the technical appendix for details of the RE and B-G IV methods

Simultaneity involving only individual effects: the Hausman-Taylor case

Model:

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

Partition $\mathbf{x}_{it}$ and $\mathbf{z}_i$:

$$\mathbf{x}_{it} = (\mathbf{x}_{1it}, \mathbf{x}_{2it}), \quad \mathbf{z}_i = (\mathbf{z}_{1i}, \mathbf{z}_{2i}),$$

where:

$$E(u_i \mid \mathbf{x}_{1it}) = 0, E(u_i \mid \mathbf{z}_{1i}) = 0 \Rightarrow \mathbf{x}_{1it}, \mathbf{z}_{1i} \text{ are exogenous}$$

$$E(u_i \mid \mathbf{x}_{2it}) \neq 0, E(u_i \mid \mathbf{z}_{2i}) \neq 0 \Rightarrow \mathbf{x}_{1it}, \mathbf{z}_{1i} \text{ are enogenous}$$

But we must assume:

$$E(\varepsilon_{it} \mid \mathbf{x}_{it}) = 0, E(\varepsilon_{it} \mid \mathbf{z}_i) = 0 \quad \text{for all x- and z-variables}$$

(no simultaneous determination of y_{it} and $(\mathbf{z}_i, \mathbf{x}_{it})$ **!!!!**)

Identification condition: no. of $\mathbf{x}_{1it} \geq$ no. of $\mathbf{z}_{2i}$

Method: use $\mathbf{x}_{1it}$ as instruments for $\mathbf{z}_{2i}$

The Hausman-Taylor (1981) estimator

Step 1: compute the within-group estimator for β:

$$\Rightarrow \text{ regress } \quad y_{it} - \bar{y}_i \text{ on } \quad \mathbf{x}_{it} - \bar{\mathbf{x}}_i \Rightarrow \quad \hat{\boldsymbol{\beta}}_W$$

Step 2: construct within-group residuals & estimate σ_ε^2 :

$$\hat{\varepsilon}_{it} = y_{it} - \bar{y}_i - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\hat{\boldsymbol{\beta}}_W$$

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{n}\sum_{t=1}^{T_i} \hat{\varepsilon}_{it}^{\,2} \Big/ \left(n(\bar{T} - 1) - k_x\right)$$

Step 3: estimate model for $\hat{e}_i = \bar{y}_i - \bar{\mathbf{x}}_i\hat{\boldsymbol{\beta}}_W$:

$$\hat{e}_i = \alpha_0 + \mathbf{z}_i\boldsymbol{\alpha} + \text{residual}, \qquad i = 1 \ldots n, \; t = 1 \ldots T_i$$

use as IVs $\mathbf{q}_{it} = [\mathbf{x}_{1it}, \mathbf{z}_{1i}]$

Step 4: Construct $\hat{e}_i^* = \bar{y}_i - \mathbf{z}_i\hat{\boldsymbol{\alpha}} - \bar{\mathbf{x}}_i\hat{\boldsymbol{\beta}}_W$; estimate σ_u^2 from $\hat{\varepsilon}_{it}$ and $\hat{e}_i^*$

Step 5: Carry out the random effects transform and estimate:

$$(y_{it} - \theta_i\bar{y}_i) = \mathbf{z}_i(1-\theta_i)\boldsymbol{\alpha} + (\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)\boldsymbol{\beta} + (1-\theta_i)u_i + (\varepsilon_{it} - \theta_i\bar{\varepsilon}_i)$$

using as IVs $\mathbf{q}_{it} = [\mathbf{z}_{1i}, (\mathbf{x}_{it} - \bar{\mathbf{x}}_i), \bar{\mathbf{x}}_{1i}]$

Endogeneity: BHPS examples

Model:

$$Ln\ wage = \alpha_0 + \alpha_1\ Female + \alpha_2\ Education\ beyond\ GCSE + \beta_1\ Age + \beta_2\ Job\ tenure + u + \varepsilon$$

(1) Is job tenure jointly determined with the wage?

- Use the standard IV/2SLS estimator in w-g form
- Possible instruments: *Married, Spouse part-time, Spouse full-time, Dissatisfied with hours,*
- But are these valid instruments?

(2) Is educational attainment influenced by the same unobservable factors as labour market success?

- Use the Hausman-Taylor estimator
- Instruments come from within the model
- But is everything uncorrelated with ε?

Within-group regression

```
. xtreg logearn age postGCSE tenure, fe

Fixed-effects (within) regression               Number of obs      =     38404
Group variable (i): pid                         Number of groups   =      7700

R-sq:  within  = 0.0983                         Obs per group: min =         1
       between = 0.0024                                        avg =       5.0
       overall = 0.0038                                        max =        11

                                                F(3,30701)         =   1115.13
corr(u_i, Xb)  = -0.4195                        Prob > F           =    0.0000

------------------------------------------------------------------------------
     logearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0249189   .0004778    52.16   0.000     .0239824    .0258554
    postGCSE |   .0263467   .0089311     2.95   0.003     .0088413     .043852
      tenure |   .0016804   .0004299     3.91   0.000     .0008377     .002523
       _cons |   .9805382   .0174738    56.11   0.000     .9462889    1.014787
-------------+----------------------------------------------------------------
     sigma_u |  .54846498
     sigma_e |  .24922759
         rho |  .82885214   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(7699, 30701) =    14.66         Prob > F = 0.0000
```

Within-group IV estimates

```
. xtivreg logearn age postGCSE (tenure = dumm*), fe
note: dumm6 dropped due to collinearity
Fixed-effects (within) IV regression          Number of obs      =         38404
Group variable: pid                           Number of groups   =          7700

R-sq:  within  = 0.0974                       Obs per group: min =             1
       between = 0.0027                                      avg =           5.0
       overall = 0.0040                                      max =            11

                                              Wald chi2(3)       =      2.40e+06
corr(u_i, Xb)  = -0.4164                      Prob > chi2        =        0.0000

------------------------------------------------------------------------------
     logearn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .0039841    .007105     0.56   0.575    -.0099415    .0179097
         age |   .0243511   .0018121    13.44   0.000     .0207995    .0279027
    postGCSE |   .0279968   .0102783     2.72   0.006     .0078518    .0481418
       _cons |   .9909042   .0363862    27.23   0.000     .9195886     1.06222
-------------+----------------------------------------------------------------
     sigma_u |  .54731645
     sigma_e |  .24934411
         rho |  .82812356   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(7699,30701) =    14.63      Prob > F    = 0.0000
------------------------------------------------------------------------------
Instrumented:   tenure
Instruments:    age postGCSE dumm1-dumm12
```

Hausman test comparing w-g regression & w-g IV

```
. hausman ivfe olsfe

                 ---- Coefficients ----
             |   (b)        (B)          (b-B)    sqrt(diag(V_b-V_B))
             |   ivfe      olsfe      Difference       S.E.
-------------+----------------------------------------------------------
      tenure | .0039841   .0016804    .0023038      .007092
         age | .0243511   .0249189   -.0005678      .001748
   postGCSE | .0279968   .0263467    .0016501      .005087
------------------------------------------------------------------------
   b = consistent under Ho and Ha; obtained from xtivreg
   B = inconsistent under Ha, efficient under Ho; obtained from xtreg

   Test:  Ho:  difference in coefficients not systematic

          chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                  =        0.11
        Prob>chi2 =      0.9912
```

⇒ No significant evidence of endogeneity in tenure

(despite the large change in the tenure coefficient when we use IV !!!)

Endogeneity of education: Hausman-Taylor

```
. xthtaylor logearn age tenure postGCSE2 female cohort, endog(tenure postGCSE2)
Hausman-Taylor estimation                       Number of obs      =     38404
Group variable (i): pid                         Number of groups   =      7700
                                                Obs per group: min =         1
                                                               avg =       5.0
                                                               max =        11
Random effects u_i ~ i.i.d.                     Wald chi2(5)       =   4111.99
                                                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
     logearn |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TVexogenous  |
         age |   .0253258   .0004155    60.95   0.000      .0245115    .0261402
TVendogenous |
      tenure |   .0016367   .0003903     4.19   0.000      .0008717    .0024016
TIexogenous  |
      female |  -.1749879   .0436307    -4.01   0.000     -.2605026   -.0894732
      cohort |   .0115968   .0033232     3.49   0.000      .0050834    .0181102
TIendogenous |
   postGCSE2 |   1.260647   .3184888     3.96   0.000      .6364202    1.884873
             |
       _cons |  -22.45571   6.338539    -3.54   0.000     -34.87902   -10.03241
-------------+----------------------------------------------------------------
     sigma_u |  1.7227596
     sigma_e |  .24925073
         rho |  .97949657   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Technical appendix 1: random effects

The following slides can be safely ignored if you're not interested in technical detail or if you aren't familiar with vector-matrix notation and matrix algebra

Random effects covariance structure

Variances & covariances (conditional on $\mathbf{z}_i$, $\mathbf{X}_i$) :

$$\text{var}(v_{it}) = \sigma_u^2 + \sigma_\varepsilon^2 ; \quad \text{cov}(v_{it}, v_{is}) = \sigma_u^2 \quad \forall\, s \neq t$$

Define the $T_i \times 1$ vector $\mathbf{v}_i$ with elements v_{i1} ... v_{iT}. Note that $\mathbf{v}_i$ and $\mathbf{v}_j$ are independent for $i \neq j$. The covariance matrix of $\mathbf{v}_i$ is:

$$\mathbf{\Omega}_i = \sigma_\varepsilon^2 \mathbf{I} + \sigma_u^2 \mathbf{E}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{E}$ is a matrix with each element equal to 1, both of order $T_i \times T_i$.

Lemma: the inverse of $\mathbf{\Omega}_i$ is:

$$\mathbf{\Omega}_i^{-1} = \frac{1}{\sigma_\varepsilon^2}\left(\mathbf{I} - \frac{T_i\sigma_u^2}{\sigma_\varepsilon^2 + T_i\sigma_u^2}\left(T_i^{-1}\mathbf{E}\right)\right) = \frac{1}{\sigma_\varepsilon^2}\left(\mathbf{M}_W + \psi_i \mathbf{M}_B\right)$$

Within- and between-group transformations

$$\mathbf{\Omega}_i^{-1} = \frac{1}{\sigma_\varepsilon^2}\left(\mathbf{M}_W + \psi_i \mathbf{M}_B\right)$$

The **M**-matrices are:

$$\mathbf{M}_W = \mathbf{I} - T_i^{-1}\mathbf{E}$$

$$\mathbf{M}_B = T_i^{-1}\mathbf{E}$$

$\mathbf{M}_W$ is the $T_i \times T_i$ idempotent matrix that transforms a $T_i \times 1$ vector of data to within-group mean deviation form;

$\mathbf{M}_B$ is the idempotent transformation to a $T_i \times 1$ vector of repeated means (the between-group transform).

The scalar $\psi_i = \sigma_\varepsilon^2 / \left(\sigma_\varepsilon^2 + T_i\sigma_u^2\right)$ reflects the relative size of $T_i\sigma_u{}^2$ and σ_ε^2 .

Generalised Least Squares

For simplicity, subsume $\mathbf{z}_i$ within $\mathbf{x}_{it}$. Then GLS is:

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Omega}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Omega}_i^{-1} \mathbf{y}_i$$

$$= \left(\sum_{i=1}^{n} \left[\mathbf{W}_{xxi} + \psi_i \mathbf{B}_{xxi} \right] \right)^{-1} \sum_{i=1}^{n} \left[\mathbf{w}_{xyi} + \psi_i \mathbf{b}_{xyi} \right]_i$$

where $\mathbf{W}_{xxi} = \sum_{t=1}^{T_i} \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i \right)' \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i \right)$, $\mathbf{B}_{xxi} = T_i \, \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i$, etc.

Maximum likelihood

If we assume u_i and ε_{it} have normal distributions, the log-likelihood function is:

$$L(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_u^2) = const - \frac{1}{2}\sum_{i=1}^{n} \ln \det \boldsymbol{\Omega}_i - \frac{1}{2}\sum_{i=1}^{n} \mathbf{v}_i' \boldsymbol{\Omega}_i^{-1} \mathbf{v}_i$$

This can be maximised numerically to estimate all parameters simultaneously.

Maximisation is done using an iterative optimisation algorithm, in which an initial guess at the parameter values is improved sequentially, until a point is reached where the gradient of the likelihood with respect to the parameters is very close to zero. Stata gives a commentary on this optimisation process.

Technical appendix 2: instrumental variables

The following slides can be safely ignored if you're not interested in technical detail or if you aren't familiar with vector-matrix notation and matrix algebra

Simultaneity: Within-group IV estimation

Model:

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

Partition $\mathbf{x}_{it}$:

$$\mathbf{x}_{it} = (\mathbf{x}_{1it}, \mathbf{x}_{2it}),$$

where: $\text{cov}(\mathbf{x}_{1it}, \varepsilon_{it}) = 0$ and $\text{cov}(\mathbf{x}_{2it}, \varepsilon_{it}) \neq 0$

Instruments $\mathbf{q}_{2it}$ (at least as many as in $\mathbf{x}_{2it}$)

where $\text{cov}(\mathbf{x}_{1it}, \varepsilon_{it}) = 0$

Full IV vector $\mathbf{q}_{it} = (\mathbf{x}_{1it}, \mathbf{q}_{2it})$

Within-group transformation:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i$$

IV estimator:

$$\hat{\boldsymbol{\beta}}_{WIV} = \left(\mathbf{W}_{xq}\mathbf{W}_{qq}^{-1}\mathbf{W}_{qx}\right)^{-1}\mathbf{W}_{xq}\mathbf{W}_{qq}^{-1}\mathbf{w}_{qy}$$

Consistency

$$\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_{WIV} = \boldsymbol{\beta} + \left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{W}_{xq}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{W}_{qq}\right)^{-1}\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{W}_{qx}\right)^{-1} \times$$

$$\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{W}_{xq}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{W}_{qq}\right)^{-1}\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{w}_{q\varepsilon}\right)$$

$$= \boldsymbol{\beta}$$

This consistency property holds because:

- The within-group transform removes u_i, which may be correlated with $\mathbf{x}_{2it}$
- The instruments are uncorrelated with ε, so:

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{w}_{q\varepsilon} = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(\mathbf{q}_{it}-\bar{\mathbf{q}}_i\right)'\left(\varepsilon_{it}-\bar{\varepsilon}_i\right) = \mathbf{0}$$

Between-group IV estimator

If $\text{cov}(\mathbf{q}_{it}, u_i) = 0$, which is a stronger requirement, then we can also use $\mathbf{q}_{it}$ as instruments in a between regression:

$$\hat{\boldsymbol{\beta}}_{BIV} = \left(\mathbf{B}_{x^*q}\mathbf{B}_{qq}^{-1}\mathbf{B}_{qx^*}\right)^{-1}\mathbf{B}_{x^*q}\mathbf{B}_{qq}^{-1}\mathbf{b}_{qy}$$

where $\mathbf{x}_{it}^{*} = \left(\mathbf{z}_i, \mathbf{x}_{it}\right)$

And then we can derive estimates of the error term variances σ_u^2 and σ_ε^2 to allow feasible GLS estimation using IV.

The random-effects IV estimator

$$\hat{\boldsymbol{\beta}}_{REIV} = \left(\mathbf{R}_{x^*q}\mathbf{R}_{qq}^{-1}\mathbf{R}_{qx^*}\right)^{-1}\mathbf{R}_{x^*q}\mathbf{R}_{qq}^{-1}\mathbf{r}_{qy}$$

where $\mathbf{R}_{x^*q} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(\mathbf{x}_{it}^* - \theta_i\bar{\mathbf{x}}_i^*\right)'\left(\mathbf{q}_{it} - \theta_i\bar{\mathbf{q}}_i\right)$, etc.

and $\theta_i = 1 - \sqrt{\sigma_\varepsilon^2 / \left(\sigma_\varepsilon^2 + T_i\sigma_u^2\right)}$

If $\mathrm{cov}(\mathbf{q}_{it}, u_i) \neq 0$, then both $\hat{\boldsymbol{\beta}}_{BIV}$ and $\hat{\boldsymbol{\beta}}_{REIV}$ are inconsistent $\Rightarrow$ a stronger requirement for instrument validity

Day 4: Binary response models

- Types of discrete variables
- Linear regression
- Latent linear regression
- Conditional (fixed-effects) logit
- Random effects logit and probit

Forms of discreteness

Censoring/corner solutions generate variables which are mixed discrete/continuous

(*e.g.* hours of work are 0 for non-employed, any positive value for employees)

Truncation involves discarding part of the population

(*e.g.* low-income targeted samples, or earnings models for employees only)

Count variables are the outcome of some counting process

(*e.g.* the number of durables owned, or the number of employees of a firm)

Binary variables reflect a distinction between two states

(*e.g.* unemployed or not, married or not)

Ordinal variables are ordered variables, possibly taking more than two values

(*e.g.* happiness on a scale 1=miserable … 5=ecstatic; rank in the army)

Unordered variables reflect outcomes which are discrete but with no natural ordering

(*e.g.* choice of occupation)

Binary models (1)

Dependent variable is

$$y_{it} = 0 \text{ or } 1$$

This describes:

- situations of choice between 2 alternatives
- sequences of events defining durations

E.g. suppose:

- $\mathbf{y}_i = (0, 0, 0, 0, 1, 1, 1, 0, 1, 1)$ is a monthly panel observation
- 0 indicates unemployment, 1 indicates employment

Then $\mathbf{y}_i$ represents a history of 4 months' unemployment followed by 3 months' employment, followed by 1 month's unemployment then 2 months' employment.

Binary models (2)

An alternative to modelling the sequence $\mathbf{y}_i$ is to model the set of durations: (U4, E3, U1, E2) $\Rightarrow$ survival analysis

An important issue concerns dynamics – how does the length of time already spent out of work affect this month's probability of finding work: *duration dependence*.

In this course, we instead focus on modelling this period's state (0 or 1):

- as a function of explanatory variables and an individual effect (static model)
- as a function of explanatory variables, an individual effect and last period's state (dynamic model). This allows for *state dependence*.

Why are special methods needed ?

Consider the binary variable, $y_{it} = 0$ or 1

Notice that the expected value of y_{it} is:

$$E(y_{it}) = \Pr(y_{it} = 1) \times 1 + \Pr(y_{it} = 0) \times 0 = \Pr(y_{it} = 1)$$

where $\Pr(y_{it} = 1)$ is the probability that $y_{it} = 1$

A simple way to model y_{it} is to use a regression with y_{it} as dependent variable. Then the RHS will be the conditional probability that $y_{it} = 1$, plus an error term.

This is called a linear probability model (LPM):

$$y_{it} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

With panel data methods (*e.g.* within-group or random-effects), the linear model implies:

$$E(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) \equiv \Pr(y_{it} = 1 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = P(\mathbf{z}_i, \mathbf{x}_{it}, u_i)$$

Disadvantages of the LPM

The linear probability model requires:

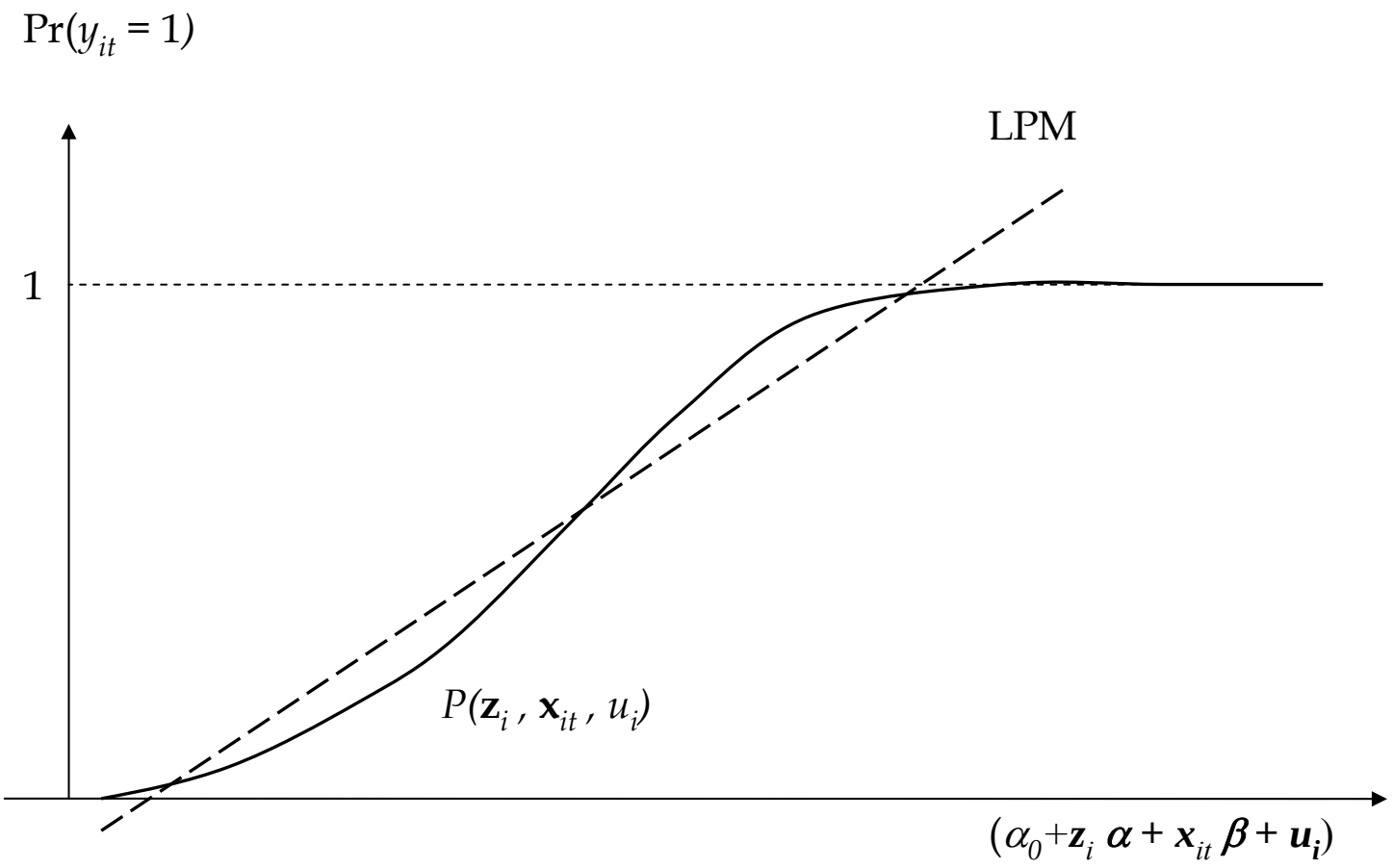
$$P(\mathbf{z}_i, \mathbf{x}_{it}, u_i) \approx \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i$$

But this may fall outside the admissible [0, 1] interval.

Moreover, $\text{var}(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = P(\mathbf{z}_i, \mathbf{x}_{it}, u_i)[1-P(\mathbf{z}_i, \mathbf{x}_{it}, u_i)]$ which varies with $\mathbf{z}_i$ and $\mathbf{x}_{it}$ $\Rightarrow$ heteroskedasticity is a problem

[Despite its disadvantages, the panel LPM is simple to estimate and is often seen in applied work – but it's not an ideal choice.]

Why nonlinear models are needed



Latent regression models: the binary case

To overcome the disadvantages of the LPM, use non-linear methods.

Define a latent (unobservable) continuous counterpart, y_{it}^*

Example from labour economics:

If y_{it}=1 defines employment, then:

y_{it}^* = best available wage – minimum acceptable wage.

Let y_{it}^* be generated by a linear regression structure:

$$y_{it}^* = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

Then employment is chosen whenever *available wage - acceptable wage* is positive:

$$y_{it} = 1 \quad \text{if and only if} \quad y_{it}^* > 0$$

Latent regression models: the binary case (2)

$$\Rightarrow \quad \Pr(y_{it} = 1 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = \Pr(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it} > 0)$$

$$= \Pr(-\varepsilon_{it} < [\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i])$$

$$= F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i)$$

where $F(.)$ is the distribution function of the random variable $-\varepsilon_{it}$

Probit model: assume ε_{it} has a normal distribution

$F(\,.\,) = \Phi(\,.\,) \Rightarrow$ df of the N(0,1) distribution

Logit (logistic regression) model: assume ε_{it} has a logistic distribution

$F(\varepsilon) = e^{\varepsilon}/[\,1+e^{\varepsilon}] \Rightarrow$ df of the logistic distribution

An aside: understanding the results from binary latent regression models

In a linear regression model:

$$y_{it} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

We can interpret the coefficients directly:

$\boldsymbol{\alpha}$ = (average) effect on y of increasing $\mathbf{z}$ by 1 unit

$\boldsymbol{\beta}$= (average) effect on y of increasing $\mathbf{x}$ by 1 unit

These are known as the *marginal effects* of $\mathbf{z}$, $\mathbf{x}$ on y

But in nonlinear models, things are more complicated. In:

$$\Pr(y_{it} = 1) = F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i)$$

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ aren't the effects on $\Pr(y_{it} = 1)$ of changing $\mathbf{z}$ or $\mathbf{x}$ by one unit $\Rightarrow$ so coefficients can't be directly interpreted

Some concepts for summarising results

Model: $\Pr(y_{it} = 1) = F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i)$

(call this conditional probability P_{it})

Coefficients	=	α_0 , $\mathbf{z}_i$ and $\boldsymbol{\beta}$
Predicted probability	=	P_{it}
Odds (O_{it})	=	$P_{it} / (1 - P_{it})$

For 2 people with different **z** and **x** –values, whose probabilities of y=1 are P_0 and P_1 :

Odds ratio	=	O_1 / O_0
Relative risk	=	P_1 / P_0

Relative risk and the odds ratio are often confused, but they are different

Marginal effects, relative risk and the odds ratio

Suppose person 0 has observable characteristics $\mathbf{z}_0$, $\mathbf{x}_0$ and unobservable characteristic u_0 ; then:

$$P_0 = F(\alpha_0 + \mathbf{z}_0 \boldsymbol{\alpha} + \mathbf{x}_0 \boldsymbol{\beta} + u_0)$$

Let's consider the effect of making a 1-unit change in (say) $\mathbf{z}$. This means inventing a new person with characteristics:

$(\mathbf{z}_0+1 , \mathbf{x}_0 , u_0)$, for whom Pr($y$=1) is:

$$P_1 = F(\alpha_0 + [\mathbf{z}_0+1]\boldsymbol{\alpha} + \mathbf{x}_0 \boldsymbol{\beta} + u_0)$$

We can summarise the effect of this change in various ways:

- *Marginal effect* $= P_1 - P_0$
- *Relative risk* $= P_1 / P_0$
- *Odds ratio* $= [P_1 / (1 - P_1)] / [P_0 / (1 - P_0)]$
  $= [P_1 / P_0] \times [(1 - P_0) / (1 - P_1)]$

Other variables are "held constant" at their baseline values$(\mathbf{x}_0 , u_0)$

Logistic regression and the odds ratio

In the logit model:

$$P_0 = exp(\alpha_0 + \mathbf{z}_0\,\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0) \;/\; [\,1 + exp(\alpha_0 + \mathbf{z}_0\,\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0)\,]$$

$$P_1 = exp(\alpha_0 + [\mathbf{z}_0 + 1]\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0) \;/\; [\,1 + exp(\alpha_0 + [\mathbf{z}_0 + 1]\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0)\,]$$

$$Odds\ ratio = [\,P_1 \;/\; (1 - P_1)\,] \;/\; [\,P_0 \;/\; (1 - P_0)\,]$$

$$= [\,exp(\alpha_0 + [\mathbf{z}_0 + 1]\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0)\,] \;/\; [\,exp(\alpha_0 + \mathbf{z}_0\,\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0)\,]$$

$$= [\,exp(\alpha_0 + \mathbf{z}_0\,\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0) \times exp(1 \times \boldsymbol{\alpha})\,] \;/\; [\,exp(\alpha_0 + \mathbf{z}_0\,\boldsymbol{\alpha} + \mathbf{x}_0\,\boldsymbol{\beta} + u_0)\,]$$

$$= exp(\boldsymbol{\alpha})$$

The odds ratio is usually only quoted in relation to logit results. It is hard to interpret and very often gets misinterpreted. It gives the proportionate effect of a 1-unit change in a variable on the ***odds*** , not the probability Pr(y=1).

Misinterpretation of odds ratios

Check that you understand the error in the following quotation:

"The odds ratio of 1.3689 for females […] indicates that, controlling for the effects of the other explanatory variables, females are 37% more likely to be in poverty than males. Stated differently, the probability of being in poverty is 1.37 times greater for females than for males."

(W. H. Crown, *Statistical Models for the Social and Behavioural Sciences: Multiple Regression and Limited Dependent Variable Models*. London: Praeger, 1998)

It isn't possible to calculate the relative risk or the marginal effect on the response probability, from knowledge of the odds ratio alone.

What would be the relative risk and marginal effect if the predicted probability for males is 0.2? What if it's 0.001? What if it's 0.8?

Options for presentation of results

- Present marginal effects evaluated at sample mean values of **x** and **z**, with individual effects *u* set at zero (*i.e.* the average in the population). But:
  - This represents a synthetic, hybrid person that doesn't exist.
  - Technically, no-one has a zero individual effect (prob is zero)
- Present *average partial effects* (APE) which allow for the average effect of the unobserved individual effects. Evaluate at:
  - Mean **x** and **z,** or
  - Selected **x** and **z** to represent typical person, or
  - Each person's **x** and **z**, and then average the results.

Other options for presentation of results

- Present predicted probabilities for different combinations of **x** and **z** (representing different types of person). Can also evaluate at different values of the individual effect u, based on its estimated distribution.
- All these methods are difficult with the fixed-effects logit, as we don't estimate the (distribution of) individual effects or the coefficients of time-invariant variables.
- Researcher should decide how to present results based on research question being asked.

Fixed effects models – some issues

- To deal with individual effects in linear FE models, we can:
  - estimate individual effects u_i (LSDV).
  - difference out individual effects u_i.

  Estimates of **β** are unaffected in both cases and are unbiased
- But in non-linear FE models:
  - There's no short-cut method of calculating the estimator without calculating the estimates of the u_i $\Rightarrow$ the "incidental parameters problem"
  - Estimated coefficients are biased
  - Can't remove the individual effects u_i by simple differencing as in within-group regression

Conditional ML estimation

- CML (as applied here) is a way of condensing the likelihood function into a form which does not depend on u_i but does depend on **β.**
- Then CML is consistent (loosely speaking, unbiased in a large sample) for **β**.
- But CML is very model specific as it is based on a technical "trick" that is only applicable in a few cases, *e.g.*:
  - logit models
  - Poisson model (for count data) – see later
- Details of conditional logit are given in the Technical Appendix

Fixed effects (or conditional) logit

Model: $\Pr(y_{it} = 1) = F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i)$,
where $F(\,.\,)$ is the logistic form

Avoiding technicalities, the method works as follows:

- Work with the subsample of individuals for whom there is some change in y_{it} during the observation period $\Rightarrow$ so we sacrifice information on any individuals displaying no change in y
- The changes in the covariates $\mathbf{x}_{it}$ (*i.e.* variable differences like $\mathbf{x}_{it}$ - $\mathbf{x}_{it}$) are then used in a modified logit analysis to explain the changes in the observed sequence of outcomes $y_{i1} \ldots y_{iT}$.
- Note that differencing the covariates removes any variables that are constant over time (*e.g.* gender, birth year, etc.), so $\boldsymbol{\alpha}$ can't be estimated
- But it also removes u_i , so we don't have to assume anything about u_i $\Rightarrow$ so FE logit is more robust than RE logit

Random effects logit/probit

Appropriate if we want to:

- estimate the coefficients of $\mathbf{z}_i$
- use a non-logistic form
- allow for dynamic adjustment (*i.e.* use the lagged value $y_{it\text{-}1}$ *as* an explanatory variable)

then conditional likelihood is not available. The random effects approach is a natural solution.

[and, of course, RE is preferred if the individual effects are independent of the **x** – use a Hausman test to decide]

Random effects logit/probit

Consider the basic model:

$$y_{it}^{*} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

$$y_{it} = 1 \quad \text{if and only if } y_{it}^{*} > 0$$

Make standard random effects assumptions (including independence of $(\mathbf{z}_i, \mathbf{x}_{it})$ and u_i).

Since the ε_{it} are independent, the joint probability of observing $(y_{i1}, y_{i1}, \ldots, y_{iTi})$ conditional on u_i (and $\mathbf{z}_i, \mathbf{x}_{it}$) is just the product of the conditional probabilities for each time period:

$$\Pr(y_{i1}, \ldots, y_{iT} \mid u_i) = \Pr(y_{i1} \mid u_i) \times \ldots \times \Pr(y_{iT} \mid u_i)$$

$$= F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{i1} \boldsymbol{\beta} + u_i) \times \ldots \times F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{iT} \boldsymbol{\beta} + u_i)$$

Random effects logit/probit

Make an assumption about the distribution of u_i (usually assumed to be N(0, σ_u^2)

Average out *(marginalise with respect to)* the unobservable u_i to get the unconditional probability of the data for individual i :

$$\Pr(y_{i1}, \ldots, y_{iT}) = \mathrm{E}[\Pr(y_{i1}, \ldots, y_{iT} \mid u_i)]$$

where "E[.]" refers to the expectation or mean with respect to the N(0, σ_u^2) distribution of u_i .

This unconditional probability $\Pr(y_{i1}, \ldots, y_{iT})$ is the likelihood for individual i. Repeated this for all individuals in the sample.

We then choose as our ML estimates the parameter values that maximise the likelihood over the whole sample. This is implemented in Stata, but computing run times are quite long.

This ML method works well only if $\mathrm{cov}(u_i, [\mathbf{z}_i, \mathbf{x}_{it}]) = 0$

Is the zero-correlation assumption valid? The Hausman test

- A Hausman test can be used to compare conditional logit estimates with the random-effects logit which assumes independence between u_i and ($\mathbf{z}_i$, $\mathbf{X}_i$).
- Null hypothesis is H_0: u_i and ($\mathbf{z}_i$, $\mathbf{X}_i$) are independent.
- Alternative hypothesis is H_1: u_i and ($\mathbf{z}_i$, $\mathbf{X}_i$) are not independent (implies we should use CL).
- $\hat{\boldsymbol{\beta}}_{CL}$ is consistent under H_0 and H_1, but inefficient under H_0 (since only uses information on changers).
- $\hat{\boldsymbol{\beta}}_{RE}$ is consistent and efficient under H_0, but inconsistent under H_1.
- Test statistic:

$$S = \left(\hat{\boldsymbol{\beta}}_{CL} - \hat{\boldsymbol{\beta}}_{RE}\right)' \left(\text{var}(\hat{\boldsymbol{\beta}}_{CL}) - \text{var}(\hat{\boldsymbol{\beta}}_{RE})\right)\left(\hat{\boldsymbol{\beta}}_{CL} - \hat{\boldsymbol{\beta}}_{RE}\right)$$

(distributed as χ^2 if H_0 is correct, with df equal to the no. of coefficients in $\boldsymbol{\beta}$)

Individual effects correlated with regressors (1)

- The RE probit/logit assumes that $(\mathbf{z}_i, \mathbf{x}_{it})$ and u_i are independent.
- Is there any way of relaxing the independence assumption?
- One possibility is to allow u_i to be correlated with elements of $\mathbf{x}_{it}$.
  - A very general formulation (due to Chamberlain) models u_i as a function of all values of $\mathbf{x}_{it}$ from all time periods.
  - A simplified version (based on the Mundlak model) is to model u_i as a function of individual means.

Individual effects correlated with regressors (2)

Using the Mundlak-style approach we have:

$$u_i = \mu + \bar{\mathbf{x}}_i\delta + \eta_i \text{ where } \eta_i \mid \bar{\mathbf{x}}_i \sim \mathrm{N}(0, \sigma_\eta^2) \qquad (1)$$

This formulation still assumes that $\mathbf{z}_i$ is not correlated with u_i. If it is, it belongs in (1), and we can't separate its correlation with u_i from its true effect. Related to this, μ absorbs the main regression constant α_0. [Can't have two constants!]

Individual effects correlated with regressors (3)

Important caveat: in linear regression, the Mundlak approximation was innocuous (the estimates of β were identical to FE). But here, we assume u_i really can be expressed as a linear function of $\bar{\mathbf{x}}_i$ such that the error term η_i is independent of $\bar{\mathbf{x}}_i$ with normal distribution.

The latent regression becomes:

$$y^*_{it} = \mu_0 + \mathbf{z}_i\boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\delta} + \eta_i + \varepsilon_{it}$$

Estimate by including individual means in list of regressors.

Unobserved heterogeneity or state dependence?

- As seen in our data set, there is much persistence in and repetition of categorical states. Past experience of a given state is often a good predictor of future experience of that state.
- Example: people who were unemployed in the past are more likely to be unemployed in the future.
- There are two possible mechanisms behind this persistence:
  - *State dependence*: experience of a given state alters behaviour in the future so as to make that state more likely to occur [see the appendix for dynamic random effects models]
  - *Unobserved heterogeneity*: individuals differ in their propensity to be in a given state and the factors explaining these differences persist over time and are unmeasured.

Technical appendix

The following slides can be safely ignored if you're not interested in technical detail or if you aren't familiar with maximum likelihood and the maths of the logit model

- Marginal effects
- Conditional logit
- Random effects likelihood function
- Dynamic random effects model

Marginal effects

- In the LPM, the marginal effect of an increase in a variable on the conditional probability that $y_{it} = 1$ is just its coefficient. Formally $\partial P(\mathbf{x}_{it}, u_i) / \partial x_{jit} = \beta_j$ (where $\mathbf{z}_i$ is absorbed into $\mathbf{x}_{it}$ for brevity)
- Note the marginal effect in the LPM does not depend on the values of other covariates, or the individual effect. So the ME is the same for everyone.
- This is not generally true in non-linear models:

$$\partial P(\mathbf{x}_{it}, u_i) / \partial x_{jit} = \partial F(\alpha_0 + \mathbf{x}_{it}\beta + u_i) / \partial x_{jit}$$
$$= f(\alpha_0 + \mathbf{x}_{it}\beta + u_i)\beta_j$$

Marginal effects (2)

- Marginal effect is coefficient multiplied by the density function (normal for probit, logistic for logit), evaluated at the base values of **x**.
- So marginal effects depend on covariates and individual effects. And usually we don't estimate the individual effects directly!
- Note we can still compare the relative effects of variables (since $f(.)$ cancels out). So the ratio of MEs due to x_j and x_k is β_j / β_k. Doesn't depend on value of latent variable.

Conditional logit

Subsume $\mathbf{z}_i$ in $\mathbf{x}_{it}$ for notational simplicity.

If we try to estimate the u_i using individual-specific dummy variables, there is no simplification analogous to within-group regression.

Moreover, the number of parameters $\rightarrow \infty$ with n, so the MLDV estimator is not consistent.

Log-likelihood for the logit model for individual i conditional on u_i:

$$L(\beta, u_1 \ldots u_n) = \sum_{t=1}^{T_i} y_{it} \ln\left(\frac{1}{1+e^{\mathbf{x}_{it}\boldsymbol{\beta}+u_i}}\right) + \sum_{t=1}^{T_i} (1-y_{it}) \ln\left(\frac{e^{\mathbf{x}_{it}\boldsymbol{\beta}+u_i}}{1+e^{\mathbf{x}_{it}\boldsymbol{\beta}+u_i}}\right)$$

The statistic $\sum_t y_{it}$ is a sufficient statistic for u_i: $\Pr(\mathbf{y}_i \mid \sum_t y_{it})$ does not depend on u_i.

Example $T_i = 2$; $\sum_t y_{it}$ can take values 0, 1, 2. Conditional on $\sum_t y_{it} = 0$, $y_{i1} = y_{i2} = 0$ and, conditional on $\sum_t y_{it} = 2$, $y_{i1} = y_{i2} = 1$ with prob 1. So only cases with $\sum_t y_{it} = 1$ are of interest.

Conditional logit (continued)

Probability of the conditioning event:

$$\Pr(\textstyle\sum_t y_{it} = 1) = \Pr(y_{i1} = 1, y_{i2} = 0) + \Pr(y_{i1} = 0, y_{i2} = 1)$$
$$= P_{i1}(1-P_{i2}) + (1-P_{i1})P_{i2}$$

$$= \frac{e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i} + e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}}{\left(1+e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}\right)\left(1+e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}\right)}$$

Conditional probability:

$$\Pr(y_{i1}=1, y_{i2}=0 \mid y_{i1}+y_{i2}=1) = \frac{\Pr(y_{i1}=1, y_{i2}=0)}{\Pr(y_{i1}+y_{i2}=1)}$$

$$= \frac{e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}}{e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i} + e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}} = \frac{e^{\mathbf{x}_{i1}\boldsymbol{\beta}}}{e^{\mathbf{x}_{i1}\boldsymbol{\beta}} + e^{\mathbf{x}_{i2}\boldsymbol{\beta}}} = \frac{e^{(\mathbf{x}_{i1}-\mathbf{x}_{i2})\boldsymbol{\beta}}}{1+e^{(\mathbf{x}_{i1}-\mathbf{x}_{i2})\boldsymbol{\beta}}}$$

$\Rightarrow$ u_i is eliminated by conditioning on $\sum_t y_{it}$

Conditional logit (continued)

With $T = 2$, the conditional log-likelihood is:

$$L(\boldsymbol{\beta}) = \sum_{i:\Sigma y=1} \left(d_i (\mathbf{x}_{i1} - \mathbf{x}_{i2})\boldsymbol{\beta} - \ln\left(1 + e^{(\mathbf{x}_{i1} - \mathbf{x}_{i2})\boldsymbol{\beta}}\right)\right)$$

where $d_i = 1$ if $y_{i1} = 1$, $y_{i2} = 0$ and 0 if $y_{i1} = 0$, $y_{i2} = 1$.

Note that, if $\mathbf{x}_{it}$ contains time-invariant covariates (*i.e.* $\mathbf{z}_i$), these disappear from $(\mathbf{x}_{i1}\text{-}\mathbf{x}_{i2}) \Rightarrow \boldsymbol{\alpha}$ cannot be estimated.

In general, conditional logit only uses data from individuals who experience change in y_{it} over time. This sacrifices sample variation.

- The same conditioning approach does not work with probit and other functional forms, nor with general dynamic models
- But it can be generalised to:
  - unordered multinomial logit models
  - ordered logit models with more than two outcomes.

The random effects likelihood function (static model)

Let $P_{it}(u_i) = \Pr(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i)$, where

$$\Pr(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = \begin{cases} F(\alpha_0 + \mathbf{z}_i\boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta} + u_i) & \text{if } y_{it} = 1 \\ 1 - F(\alpha_0 + \mathbf{z}_i\boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta} + u_i) & \text{if } y_{it} = 0 \end{cases}$$

Then the likelihood function for individual i, conditional on u_i, *is* :

$$L_i(u_i) = \prod_{t=1}^{T} P_{it}(u_i) \quad ,$$

which tells us, for given values of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ_u^2 and σ_ε^2, and given value of u_i how well the model fits the data on individual i.

Integrating out the random effects

Including u_i in the conditioning set greatly simplifies the likelihood function, because errors from different time periods are then independent (otherwise, we'd need to allow for dependence across periods).

But… we don't know u_i (also we have the incidental parameters problem). We do, however, know (by assumption!) its distribution. Therefore we can "average out " or marginalise with respect to u_i:

$$L_i = E\left(\prod_{t=1}^{T_i} P_{it}(u_i)\right) = \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} P_{it}(u) g(u) du$$

where $g(u)$ is an assumed density for u, e.g. for probit, Gaussian: $g(u) = \sigma_u^{-1}\phi(u/\sigma_u)$. The full likelihood function is $L = \Pi L_i$

Evaluation of the likelihood function requires the integral to be approximated numerically by a quadrature algorithm.

Day 5: Further topics

- Ordered response models
- Incomplete panels and sample selection in panel data models
- Dynamic fixed-effects regression models
- Dynamic binary logit/probit models
- Policy evaluation and panel data
- Count data models

Topic 1:

Ordered response models

Ordered response models

- Ordered (or ordinal) variables take discrete values which have a natural ordering:
  - Happiness on a scale of 1-5
  - Not working, part-time, full-time
  - Want fewer, same, more work hours
  - No, part, full insurance
  - Credit rating
- Variables are ordinal but not (necessarily) cardinal, i.e. the "distance" between two categories has no meaning in the model. Only order matters.

Latent regression (1)

- As in binary response models, assume there is an underlying latent variable y_{it}^* determined as follows:

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

- u_i is assumed to be a random effect distributed independently of $(\mathbf{z}_i, \mathbf{X}_i)$ as $N(0, \sigma_u^2)$.
- Note there is no constant (see later).
- The observed value of y_{it} is $\{0, 1, \ldots, J\}$, depending on where y_{it}^* falls relative to a set of J *cutpoints* or *thresholds*, $\mu_1 < \mu_2 < \ldots < \mu_J$.

Latent regression (2)

- The outcome y_{it} is given as:

  $y_{it} = 0$ if $y_{it}^* \leq \mu_1$

  $y_{it} = 1$ if $\mu_1 < y_{it}^* \leq \mu_2$

  .

  $y_{it} = J$ if $\mu_J < y_{it}^*$

- So, if $J = 3$, there are 2 cutpoints, μ_1 and μ_2.
- And if $J = 2$ (binary choice model), there is only one cutpoint, μ_1.
  - This is slightly different to the usual specification of the binary probit/logit. Usually, μ_1 is normalised to zero and a constant included in the list of regressors. Here, we set the constant to zero and estimate μ_1, as is done in Stata's `oprobit` and `reoprob`. The choice is arbitrary.

Random effects ordered probit (1)

- Assume ε_{it} is normally distributed with unit variance.

$$\Pr(y_{it} = 0 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = \Pr(y_{it}^* \leq \mu_1 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i)$$
$$= \Pr(\mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + u_i + \varepsilon_{it} \leq \mu_1)$$
$$= \Phi(\mu_1 - \mathbf{z}_i \alpha - \mathbf{x}_{it} \beta - u_i)$$

$$\Pr(y_{it} = 1 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = \Pr(\mu_1 < y_{it}^* \leq \mu_2 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i)$$
$$= \Pr(\mu_1 < \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + u_i + \varepsilon_{it} \leq \mu_2)$$
$$= \Phi(\mu_2 - \mathbf{z}_i \alpha - \mathbf{x}_{it} \beta - u_i) - \Phi(\mu_1 - \mathbf{z}_i \alpha - \mathbf{x}_{it} \beta - u_i)$$

[which is just $\Pr(y_{it}^* \leq \mu_2)$ minus $\Pr(y_{it}^* \leq \mu_1)$]

Etc…

Random effects ordered probit (2)

- Finally:

$$\Pr(y_{it} = J \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = \Pr(\mu_J < y_{it}^* \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i)$$
$$= 1 - \Pr(y_{it}^* \leq \mu_J \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i)$$
$$= 1 - \Phi(\mu_J - \mathbf{z}_i\alpha - \mathbf{x}_{it}\beta - u_i)$$

- Check that these probabilities sum to one!
- Predicting probabilities and calculating marginal effects is done analogously to the binary RE probit.
- But there is a complication in the intermediate categories 1, … J.

Marginal effects

- For example (absorb $\mathbf{z}_i$ into $\mathbf{x}_{it}$ for brevity):
  $\Pr(y_{it}=1 \mid \mathbf{x}_{it}, u_i) = \Phi(\mu_2 - \mathbf{x}_{it}\beta - u_i) - \Phi(\mu_1 - \mathbf{x}_{it}\beta - u_i)$

- So the marginal effect of $\mathbf{x}_{jit}$ on the probability that $y_{it}=1$ is:

$\partial \Pr(y_{it}=1 \mid \mathbf{x}_{it}, u_i)/\partial x_{jit} = -\beta_j\phi(\mu_2 - \mathbf{x}_{it}\beta - u_i) + \beta_j\phi(\mu_1 - \mathbf{x}_{it}\beta - u_i)$

- This can be either negative or positive (consider the $\phi(.)$ function). And in general, the sign will vary with $\mathbf{x}_{it}$ and u_i.
  - Intuitively, why does the marginal effect have an ambiguous sign?

Topic 2:

Incomplete panels and sample selection in panel data models

Incomplete panels

- We have distinguished between balanced, unbalanced and non-compact panels.
- Most techniques (Stata commands) can be used with all three types of panel.
- But…
  - We have implicitly assumed that missing observations only represent an efficiency loss (i.e. estimates are still unbiased).
  - In fact, the pattern of missing observations may not be random.
  - If observations are not missing at random, estimates may be biased. Thus unbalanced and non-compact panels may not be random samples.
  - Equally, balanced (sub-)panels may not be random – respondents present at every wave are unlikely to be representative of the population.

Non-response

- Why might observations be missing?
- Unit non-response
  - Attrition – respondents drop out of panel
  - Wave non-response - unavailable at particular waves
- Item non-response
  - Respondents fail to answer particular questions, e.g. income.
- Types of missing-ness:
  - Missing completely at random (MCAR)
  - Missing at random (MAR): conditional on observables ($\mathbf{X}_i$, $\mathbf{z}_i$), response is random. Systematic differences in response are explained by observable characteristics.
  - Informative or non-ignorable non-response: systematic differences in response remain after controlling for ($\mathbf{X}_i$, $\mathbf{z}_i$).

Implications of incompleteness

- Implications depend on type of analysis (but this is a complex area with disagreements between econometricians and survey statisticians).
- Descriptive (i.e. unconditional) statistics will be unbiased if data are MCAR, but biased if data are MAR or non-response is informative.
  - Example: if poor households are less likely to participate in surveys, we will underestimate the poverty rate.
- Conditional estimates (regressions) are unbiased if data are MCAR or MAR (conditional on observables in model). Biased if non-response is informative.

Weights?

- Data sets usually include weights which account for:
  - systematic non-response (as a function of particular observables);
  - non-representative sampling due to survey design.
- Use weights for descriptive stats (if want to make inferences about the population).
- Weighting is more problematic in regression analysis:
  - General purpose weighting model may not be appropriate for a specific regression model
  - May be identification problems if same variables used for weights and in regression.
  - Weighting is not necessary if data are MAR, and inflates SEs.
  - In practice, Stata does not accept weights for linear FE and RE (GLS) analysis.

Non-random selection in panels

- In the regression framework, non-random response can be represented as follows. Let the model of interest be:

  $y_{it} = \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}, \quad t = 1 \ldots T, i = 1 \ldots n$
- Define a response indicator r_{it} which equals 1 if (y_{it}, $\mathbf{z}_i$, $\mathbf{x}_{it}$) is observed in the panel and 0 otherwise.
- If data are MCAR or MAR, then r_{it} is independent of u_i and ε_{it}.
- If non-response is non-ignorable then r_{it} is not independent of u_i and ε_{it}. Also called non-random selection or selection on unobservables.

Consequences for RE estimates

- We focus on the implications of missing observations for linear RE and FE estimates.
- RE is unbiased if:

  $$E(u_i + \varepsilon_{it} \mid \mathbf{X}_i, \mathbf{z}_i, r_i) = E(u_i + \varepsilon_{it} \mid \mathbf{X}_i, \mathbf{z}_i) = 0$$

  where $r_i = (r_{i1}, \ldots, r_{iT})$, a vector of selection outcomes in all periods.

  This says that the composite error term is unrelated to selection conditioning on observable characteristics (MAR or selection on observables).

Consequences for FE estimates

- Unsurprisingly (why?), FE is more robust to non-random selection into the panel.
- FE is consistent if:

  $$E(\varepsilon_{it} \mid \mathbf{X}_i, \mathbf{z}_i, u_i, r_i) = E(\varepsilon_{it} \mid \mathbf{X}_i, \mathbf{z}_i, u_i) = 0$$

  This says that the transitory error term is unrelated to selection, conditioning on observable characteristics and the individual effect u_i. But r_i *can* be related to u_i.
- As long as selection into the panel works through "levels", i.e. time-invariant factors, then FE remains consistent.

Testing for non-random selection in panels

- Some simple indicative tests for non-random selection involve:
  1. checking whether r_i helps explain the outcome y_{it} after controlling for other characteristics
  2. comparing results from the unbalanced panel with the balanced sub-panel.

- In the first type of test, functions of r_{it} can be added to the equation and their significance tested [note r_{it} can't be added – why not?]. For example:
  - lagged response indicator $r_{it\text{-}1}$
  - indicator for present in all waves, $c_i = \Pi\, r_{it}$
  - number of waves present for, $T_i = \Sigma\, r_{it}$

  The last two can only be used with RE (why?).

"Hausman" test

2. A second test compares RE or FE estimates from the unbalanced panel and its balanced sub-panel. If selection is random, the two estimates should be close. If selection is non-random, and affects the estimators differently, we expect a statistically significant difference between the two.

   For example, test the RE estimator by forming the statistic:

$$\left(\hat{\boldsymbol{\beta}}_{RE,B} - \hat{\boldsymbol{\beta}}_{RE,U}\right)' \left[\operatorname{var}\left(\hat{\boldsymbol{\beta}}_{RE,B}\right) - \operatorname{var}\left(\hat{\boldsymbol{\beta}}_{RE,U}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}}_{RE,B} - \hat{\boldsymbol{\beta}}_{RE,U}\right)$$

$$\sim \chi^2(k) \text{ under } H_0 : \text{no selection bias}$$

   [Not a true Hausman test because neither estimator is consistent in presence of selection bias, and both $\hat{\boldsymbol{\beta}}_{RE,U}$ and $\hat{\boldsymbol{\beta}}_{RE,B}$ may be affected similarly by selection. Thus the test may have low "power"]

If these tests suggest attrition bias, the situation is difficult: methods to correct for "endogenous" attrition are complicated

Topic 3:

Dynamic fixed-effects regression models

Dynamic models

Why model dynamics?

- Current outcomes might depend on past values of determinants ➔ include lagged **x**s (distributed lag model). Use similar techniques to those already discussed.
- Adjustment might be partial: this year's outcome y depends not only on **x,** but also on last year's outcome ➔ include lagged y. We will focus on this case.
  - Notice (as we will see) that this amounts to including an infinite (or back to start of process) number of lagged **x**.

Dynamic models for continuous dependent variables

Adjustment may be imperfect – how to model it? Any conventional time-series model can be used, *e.g.* AR(1):

$$y_{it} = \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma y_{it-1} + u_i + \varepsilon_{it} \qquad (1)$$

or static model with AR(1) errors:

$$y_{it} = \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + u_i + \varepsilon_{it} \qquad (2)$$

$$\varepsilon_{it} = \rho \varepsilon_{it-1} + \eta_{it}$$

$$\Rightarrow \quad y_{it} = \mathbf{z}_i (1-\rho)\alpha + (\mathbf{x}_{it} - \rho \mathbf{x}_{it-1})\beta + \rho y_{it-1} + u_i + \eta_{it} \qquad (2')$$

NB: model (1) implies gradual adjustment to change in **x**; model (2) implies a full immediate response.

More general distributed lag models can be used (e.g. ECMs, ARMA, etc.)

Within-group estimation

Within-group transformed model (e.g. AR(1)):

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + \gamma(y_{it-1} - \bar{y}_i^*) + \varepsilon_{it} - \bar{\varepsilon}_i$$

where:

$$\bar{y}_i^* = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it-1} = \frac{1}{T_i}\sum_{t=0}^{T_i-1} y_{it} \neq \bar{y}_i$$

NB we assume a compact panel (why?) and an observable initial condition y_{i0}

We have got rid of the individual effect. But what are the statistical properties of a regression of

$y_{it} - \bar{y}_i$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ and $(y_{it-1} - \bar{y}_i^*)$?

Properties of the within-group estimator (1)

Find an expression for y_{it} that only involves $\mathbf{z}$, $\mathbf{x}$, and y_{i0} (the starting value or "initial condition" of y).

$$y_{it} = \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma y_{it-1} + u_i + \varepsilon_{it}$$

By substitution:

$$\begin{aligned}
y_{it} &= \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma (\mathbf{z}_i \alpha + \mathbf{x}_{it-1} \beta + \gamma y_{it-2} + u_i + \varepsilon_{it-1}) + u_i + \varepsilon_{it} \\
&= (1+\gamma)\, \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma \mathbf{x}_{it-1} \beta + \gamma^2 y_{it-2} + u_i + \gamma u_i + \gamma \varepsilon_{it-1} + \varepsilon_{it} \\
&= (1+\gamma)\, \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma \mathbf{x}_{it-1} \beta + \gamma^2 (\mathbf{z}_i \alpha + \mathbf{x}_{it-2} \beta + \gamma y_{it-3} + u_i + \varepsilon_{it-2}) \\
&\quad + u_i + \gamma u_i + \gamma \varepsilon_{it-1} + \varepsilon_{it} \\
&= (1+\gamma+\gamma^2)\, \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma \mathbf{x}_{it-1} \beta + \gamma^2 \mathbf{x}_{it-2} \beta + \gamma^3 y_{it-3} \\
&\quad + u_i + \gamma u_i + \gamma^2 u_i + \varepsilon_{it} + \gamma \varepsilon_{it-1} + \gamma^2 \varepsilon_{it-2}
\end{aligned}$$

And so on… Eventually we arrive at $t=0$.

Properties of the within-group estimator (2)

Distributed lag form of (1):

$$y_{it} = \sum_{s=0}^{t-1} \gamma^s \left(\mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it-s} \boldsymbol{\beta} + u_i + \varepsilon_{t-s} \right) + \gamma^t y_{i0}$$

$$= \frac{1-\gamma^t}{1-\gamma} \left(\mathbf{z}_i \boldsymbol{\alpha} + u_i \right) + \sum_s \gamma^s \mathbf{x}_{it-s} \boldsymbol{\beta} + \left[\varepsilon_{it} + \gamma \varepsilon_{it-1} + \ldots + \gamma^{t-1} \varepsilon_{i1} \right] + \gamma^t y_{i0}$$

$\Rightarrow$ $y_{it\text{-}1}$ is a function of $\varepsilon_{it\text{-}1} \ldots \varepsilon_{i1}$

$\Rightarrow$ $\bar{y}_i^* = \sum_{t=0}^{T_i - 1} y_{it} \Big/ T_i$ is a function of $\varepsilon_{iT\text{-}1} \ldots \varepsilon_{i1}$ and y_{i0}

$\Rightarrow$ $y_{it-1} - \bar{y}_i^*$ is correlated with $\varepsilon_{it} - \bar{\varepsilon}_i$

$\Rightarrow$ bias in within-group regression coefficients

Properties of the within-group estimator (3)

- Bias of the within-groups estimator is caused by eliminating the individual effect u_i from the equation. This causes a correlation between the transformed error term and the transformed lagged dep var.
- Bias is generally *negative* for small T (even if true γ is zero).
- For large T, bias is small – but with panel data T is not usually large…

What about pooled OLS?

Properties of the pooled OLS estimator

- Assume individual effects u_i are random. In a static model, OLS is unbiased and consistent (though, recall, inefficient).
- But this is not the case in a dynamic model:

  $y_{it} = \mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + \gamma y_{it-1} + u_i + \varepsilon_{it}$
- We know from above that y_{it-1} is a function of u_i and y_{i0}. In general, correlation between y_{it-1} and $u_i + \varepsilon_{it}$ is positive due to:
  - Positive contribution from u_i.
  - Positive contribution from y_{i0} if y_{i0} generated by same process as any other y_{it}
- So OLS is biased upward and is inconsistent

Other estimators?

- GLS and ML estimators are also generally biased
  - They depend critically on assumptions about initial conditions y_{i0}, and how they are generated
- There are several IV estimators which correct for endogeneity of the lagged dependent variable and are also independent of initial conditions. Like HT, instruments come from inside the model.
  - Anderson-Hsiao
  - Arellano-Bond
  - Blundell-Bond
  - …

A simple IV estimator

The within-group transform complicates estimation with lagged endogenous variables. Consider time-differencing:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \gamma \Delta y_{it-1} + \Delta \varepsilon_{it} \ , \qquad t = 2 \ldots T_i \qquad (1)$$

The problem now is that the error term, $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it\text{-}1}$ is a MA(1) process which contains $\varepsilon_{it\text{-}1}$, which is correlated with $\Delta y_{it\text{-}1}$.

$\Rightarrow$ Find a set of instruments correlated with $\Delta y_{it\text{-}1}$ but uncorrelated with $\varepsilon_{it\text{-}1}$

$\Rightarrow$ All lagged $\mathbf{x}_{it}$ and $y_{it\text{-}2} \ldots y_{i0}$ are valid instruments if $\{\varepsilon_{it}\}$ is serially independent

$\Rightarrow$ Simplest IV estimator (Anderson Hsiao) estimates (1), using instruments $(\mathbf{x}_{it}, \mathbf{x}_{it\text{-}1}, \mathbf{x}_{it\text{-}2}, y_{it\text{-}2})$.

$\Rightarrow$ We can only use observations $t = 2 \ldots T_i$. Each extra lag used as an instrument loses us n observations.

$\Rightarrow$ Once $\hat{\boldsymbol{\beta}}_{IV}$ is found, estimate $\boldsymbol{\alpha}$ by regressing $\bar{y}_i - \bar{\mathbf{x}}_i \hat{\boldsymbol{\beta}}_{IV}$ on $\mathbf{z}_i$

Problems with IV estimators

Suppose y_{it} is a random walk (e.g. Hall's (1978) form of the permanent income hypothesis: dynamic choice models based on Euler conditions).

$\Rightarrow y_{it\text{-}2}$ is uncorrelated with $\Delta y_{it\text{-}1}$ and is not a valid instrument

$\Rightarrow$IV methods based on a differenced model won't work well if there is a near-unit root

Any method based solely on the differenced equation ignores potentially valuable information contained in the initial condition y_{i0}

What is the optimal point on the trade-off between the number of lags used as instruments and the number of time periods retained in the estimation sample?

System estimators

The time-differenced model:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \gamma \Delta y_{it-1} + u_i + \Delta \varepsilon_{it} \ , \qquad t = 2 \ldots T_i \qquad (1)$$

This is a system of T_i-1 linear equations with cross-correlated errors (since $\Delta\varepsilon_{it}$ is correlated with $\Delta\varepsilon_{it\text{-}1}$ and $\Delta\varepsilon_{it+1}$)

There is also some (related) process generating the initial conditions, y_{i0} and y_{i1}, which could provide further equations.

A different number of instruments is available for each of the equations in (1):

E.g. the equation for $t = 2$ has only $(\mathbf{x}_{i0} \ldots \mathbf{x}_{iT}, y_{i0})$;

the equation for $t = T_i$ has $(\mathbf{x}_{i0} \ldots \mathbf{x}_{iT}, y_{i0} \ldots y_{iT\text{-}2})$.

NB it's assumed here that $\mathbf{x}_{i0}$ is observable

Digression: method of moments (1)

The method of moments is a way of getting consistent estimates of model parameters.

1. Specify moment conditions (e.g. means, covariances) implied by the model as a function of its parameters (population moments).
2. Write down the "sample analogues" of these moment conditions, i.e. expressions into which you can plug the sample data, as a function of parameter estimates.
3. Choose values for the parameter estimates which "solve" the sample moment conditions.

Digression: method of moments (2)

Very simple example: mean of a random variable y.

1. Mean of y is defined as $\mu = E[y]$. Rearrange as a moment condition: $m(y; \mu) = E[y - \mu] = 0$.

2. Sample analogue is $\hat{m}(\mathbf{y}; \mu) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu) = 0$

3. Solve to get MM estimator: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Digression: method of moments (3)

- Often there are more moment conditions than parameters to be estimated. Then the moment conditions don't have a unique solution.
- In this case, we minimise a (weighted) sum of the squares of the sample moments. In vector notation this is written in the general case as $\hat{\mathbf{m}}(\mathbf{y},\mathbf{x},\boldsymbol{\beta})'\mathbf{V}^{-1}\hat{\mathbf{m}}(\mathbf{y},\mathbf{x},\boldsymbol{\beta})$ where V is the weighting matrix.
- This is called the generalised method of moments (GMM).

Generalised method of moments

IV estimators are members of the class of GMM estimators

e.g. the 2SLS estimator, $\hat{\boldsymbol{\beta}}_{IV} = \left(\mathbf{X}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}$

is the following M-estimator:

$$\hat{\boldsymbol{\beta}}_{IV} = \arg\min_{\boldsymbol{\beta}} \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)$$

$$= \arg\min_{\boldsymbol{\beta}} \hat{\mathbf{m}}(\mathbf{y},\mathbf{x},\boldsymbol{\beta})'\mathbf{V}^{-1}\hat{\mathbf{m}}(\mathbf{y},\mathbf{x},\boldsymbol{\beta})$$

where $\hat{\mathbf{m}}$ is the "sample analogue", $n^{-1}\mathbf{Q}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, of a moment, $E\mathbf{q}'\varepsilon$, assumed to be zero in the population.

$\mathbf{V}$ is a weighting matrix proportional to the asymptotic covariance matrix of the moment condition (in this standard 2SLS example $\sigma_\varepsilon^2\mathbf{Q}'\mathbf{Q}$, where σ_ε^2 is the residual variance).

GMM can be extended to any number of moment conditions

Arellano-Bond GMM (1991)

We have T_i -2 differenced equations (1).

The instruments for equation t are:

$$\mathbf{q}_{it} = (\mathbf{x}_{i0} \ldots \mathbf{x}_{iT}, y_{i0} \ldots y_{it-2})$$

Full set of moment conditions:

$$E\, \mathbf{q}_{i2}' \Delta\varepsilon_{i2} = 0 \qquad (T_i+1)k_x+1 \text{ conditions}$$

$$E\, \mathbf{q}_{i3}' \Delta\varepsilon_{i3} = 0 \qquad (T_i+1)k_x+2 \text{ conditions}$$

$$\vdots$$

$$E\, \mathbf{q}_{iT}' \Delta\varepsilon_{iT} = 0 \qquad (T_i+1)k_x+T_i-1 \text{ conditions}$$

$\hat{\mathbf{m}}$ is a $[(T_i+1)(T_i-1)k_x + T_i(T_i-1)/2] \times 1$ moment vector

The optimal choice for $\mathbf{V}$ is $E\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i'$

More conditions can be added (*e.g.* for $\mathbf{z}_i$ and to impose the homoskedasticity assumption on ε_{it}). ***But*** GMM often works badly in finite samples with many moment conditions.

Specification testing
(1) Testing for over-identifying restrictions

The number of restrictions = the number of moment conditions for each individual (r) *minus* the number of parameters (k_x).

Sargan test statistic:

The minimized optimal GMM criterion scaled by n is

$$S = n\left(\hat{\mathbf{m}}(\mathbf{y},\mathbf{x},\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}\hat{\mathbf{m}}(\mathbf{y},\mathbf{x},\hat{\boldsymbol{\beta}})\right)$$

has an asymptotic chi-square distribution with r- k_x degrees of freedom.

Specification testing
(2) Testing for residual serial correlation

If the ε_{it} are serially independent, then

$E[\Delta \varepsilon_{it} \Delta \varepsilon_{it-1}]=E[(\varepsilon_{it} - \varepsilon_{it-1})(\varepsilon_{it-1} - \varepsilon_{it-2})]=-E[\varepsilon_{it-1}^2]=-\sigma_\varepsilon^2$

Also $var(\varepsilon_{it} - \varepsilon_{it-1})=var(\varepsilon_{it-1} - \varepsilon_{it-2})=2\sigma_\varepsilon^2$

Thus, the first order serial correlation coefficient is

$r_1=E[\Delta \varepsilon_{it} \Delta \varepsilon_{it-1}]/[\sqrt{var(\Delta\varepsilon_{it})}\sqrt{var\Delta \varepsilon_{it-1}}] = 0.5.$

But $E[\Delta \varepsilon_{it} \Delta \varepsilon_{it-2}]=0$, and so the second order serial correlation coefficient r2=0.

$\Rightarrow$ test for second order serial correlation.

Specification error if second order serial correlation is statistically significant.

Further developments: initial conditions

Arellano-Bond ignores the initial conditions y_{i0} and y_{i1} and only uses moment conditions for $\Delta y_{i2} \ldots \Delta y_{iT}$.

To progress further, we need additional assumptions about the initial conditions. One possibility is:

Equilibrium initial values. If the process is homogeneous and long-established:

$$y_{i0} = \frac{\mathbf{z}_i \boldsymbol{\alpha} + u_i}{1-\gamma} + \sum_{s=0}^{\infty} \gamma^s \left(\mathbf{x}_{i,-s} \boldsymbol{\beta} + \varepsilon_{i,-s} \right)$$

$\Rightarrow$Coefficient of u_i in equation for y_{i0} is $(1-\gamma)^{-1}$

$\Rightarrow$But the quantity $\sum_{s=0}^{\infty} \gamma^s \mathbf{x}_{i,-s}$ is unobserved

$\Rightarrow$Also, do people really have infinite pasts?

If lagged levels of y_{it} are poor instruments for Δy_{it-1}, can we go back to using the equations in level form?

Extended system methods

Arellano & Bover (1995) and Blundell & Bond (1998) (see also Bhargava & Sargan, 1983) suggested using the model in *both* differenced and levels form to generate GMM moment conditions.

Question: in the levels model

$$y_{it} = \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it-1} + u_i + \varepsilon_{it} \ , \qquad (1)$$

is there a good instrument for y_{it-1}? This instrument must be uncorrelated with u_i as well as ε_{it} .

A&B suggested Δy_{it-1} , *etc*. The instrument validity condition is $E[\Delta y_{it-1} (u_i + \varepsilon_{it})] = 0$, which requires (see B&B, 1998):

$$E\, u_i [\, y_{i0} - u_i/(1-\gamma)] = 0 \qquad (2)$$

$$E\, u_i \Delta \varepsilon_{it} = 0 \qquad (3)$$

(2) Requires y_{i0} to be in stationary equilibrium. It then improves estimation precision in highly-persistent models (*i.e.* when $\gamma \approx 1$)

Topic 4:

Dynamic binary logit/probit models

Dynamic binary models

- Unobserved (time-invariant) heterogeneity will lead to persistence over time after controlling for all observable characteristics, even if there is no true state dependence.
- We often want to measure, or control for, true state dependence, e.g. does past experience of unemployment make future unemployment more likely? Implies long term effects of econ policy.
- Dynamic models using panel data allow both unobserved heterogeneity and state dependence.

Dynamic random effects binary models

- We focus on the RE binary model (logit or probit) with a simple dynamic specification (one lag of the dependent variable).
- The latent regression is now:

  $y_{it}^{*} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it-1} + u_i + \varepsilon_{it}$

  $y_{it} = 1$ if and only if $y_{it}^{*} > 0$
- True state dependence is measured by γ, and persistent unobserved heterogeneity is captured by u_i
- Assume (as previously) that ε_{it} is serially uncorrelated

The random effects likelihood function

Construct a likelihood by sequential conditioning:

$$\Pr(y_{i0} \mid \mathbf{z}_i , \mathbf{X}_i , u_i) = P_{i0}(u_i)$$
$$\Pr(y_{i1} \mid y_{i0} , \mathbf{z}_i , \mathbf{x}_{i1} , u_i) = P_{i1}(y_{i0} , u_i)$$
$$\vdots$$
$$\Pr(y_{iT} \mid y_{iT-1} , \mathbf{z}_i , \mathbf{x}_{iT} , u_i) = P_{iT}(y_{iT-1} , u_i)$$

The probabilities P_{it} (for $t = 1, \ldots, T$) are of the form:

$$F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it-1} + u_i) \quad \text{for } y_{it} = 1$$

or $$1 - F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it-1} + u_i) \quad \text{for } y_{it} = 0.$$

Likelihood function for individual i, conditional on u_i:

$$L_i(u_i) = P_{i0}(u_i) \prod_{t=1}^{T_i} P_{it}(y_{it-1}, u_i)$$

Initial conditions

- The $P_{i0}(u_i)$ term in the likelihood is the contribution of the initial condition – the first observed value y.
- If y_{i0} is exogenous (unrelated to the individual effect) then effectively $P_{i0}(u_i)$ can be dropped from the likelihood
  - Just condition on y_{i0} in $P_{i1}(y_{i0}, u_i)$
  - Possible efficiency loss since useful information about the starting point may be neglected.
- But y_{i0} is probably not exogenous:
  - It is probably not the true starting point of the "process", just the start of our sample
  - In any case, y_{i0} is probably not randomly allocated, but related to u_i as are the other y_{it}.

Heckman's method

- In practice, it is difficult to derive an exact expression for $P_{i0}(u_i)$, especially if we do not observe the process from the beginning.
- Heckman (1981) suggested approximating $P_{i0}(u_i)$ by a simple probit model, where regressors can include "pre-sample" information (e.g. family background).
- Can be complicated to estimate.

Wooldridge's method

Wooldridge suggested an alternative: condition on y_{i0}, without specifying its probability. Instead, model the density of u_i conditional on y_{i0}, $\mathbf{x}_i$. This is related to the Chamberlain/Mundlak approach discussed earlier.

So u_i could be specified as:

$$u_i = \mu + \bar{\mathbf{x}}_i\delta + \gamma_0 y_{i0} + \eta_i \text{ where } \eta_i \mid \bar{\mathbf{x}}_i, y_{i0} \text{ is distributed as } \mathrm{N}(0, \sigma_\eta^2)$$

and the latent regression is now:

$$y_i^* = \mu_0 + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\alpha} + \gamma y_{it-1} + \bar{\mathbf{x}}_i\delta + \gamma_0 y_{i0} + \eta_i + \varepsilon_{it}$$

Can be estimated as standard RE probit – include $\bar{\mathbf{x}}_i$ and y_{i0} every period.

Again, though, note this is just an approximation.

Topic 5:

Policy evaluation and panel data

Policy evaluation and panel data

- A specialised application of statistics is to evaluate the impact of various new policies, e.g. training schemes, changes to tax-benefit system, minimum wages.
- Policy evaluation often uses panel data.
- We look briefly at the parameters that policy evaluation methods try to measure and how they relate to panel data estimators seen earlier in the course.

Potential outcomes and counterfactuals

- Aim is to evaluate impact of some policy 'treatment' (terminology originates in clinical trials).
- Each individual has two potential outcomes, y_{1i} (with treatment) and y_{0i} (without treatment).
- The treatment effect is $\Delta_i = y_{i1} - y_{i0}$. Note that Δ_i potentially differs over individuals (e.g. some people benefit more from training than others).
- Problem is we only observe each individual in one state (treated or untreated). We don't observe the counterfactual state, i.e. what would have happened to the treated person had they not been treated, and the untreated person had they been treated.

Average treatment effects (1)

- Say we want to estimate the average effect of the treatment. The *population average treatment effect* (ATE) is $E(\Delta_i) = E(y_{1i} - y_{0i}) = E(y_{1i}) - E(y_{0i})$. But, as already seen, we don't observe y_{1i} and y_{0i} for all individuals in the sample.
- But, using available observations, we could estimate (naively): $E(y_{1i} \mid d_i = 1) - E(y_{0i} \mid d_i = 0)$

$= E(y_{1i} \mid d_i = 1) - E(y_{0i} \mid d_i = 1) + E(y_{0i} \mid d_i = 1) - E(y_{0i} \mid d_i = 0)$

$= E(y_{1i} - y_{0i} \mid d_i = 1) + E(y_{0i} \mid d_i = 1) - E(y_{0i} \mid d_i = 0)$

$= \text{ATT} + E(y_{0i} \mid d_i = 1) - E(y_{0i} \mid d_i = 0)$

where d_i indicates treatment and ATT is the *average effect of treatment on the treated*.

Average treatment effects (2)

- ATT will often differ from ATE. E.g. training may be given to those who benefit the most from it. But ATT is often the more relevant parameter for policy purposes – e.g. want to know the impact on those who will actually participate in a scheme.
- The naïve estimator includes a bias/selection term $E(y_{0i} \mid d_i = 1) - E(y_{0i} \mid d_i = 0)$, which is the difference in untreated outcomes between those who got the treatment and those who didn't. This term will not be zero if, e.g., trainees would have earned less (or more) than non-trainees even without training.

Before-after estimator (1)

The bias term highlights the key problem in policy evaluation, which is making sure that the treated and untreated groups are very similar (ideally, identical). On average, the outcomes of the 2 groups should be the same in the absence of the treatment.

Consider a possible estimator using two waves of panel data (*t* and *t*+1), with treatment occurring after the first wave. Compare treated individuals with their "untreated selves" in the previous wave, i.e. estimate:

$$E(y_{1it+1} \mid d_i = 1) - E(y_{0it} \mid d_i = 1)$$

by $\bar{y}^T_{t+1} - \bar{y}^T_t$, where $\bar{y}^T$ is the mean outcome for treated individuals

Before-after estimator (2)

- The before-after estimator uses outcomes before treatment (at t) to proxy (non-observed) outcomes at t+1 without the treatment. It identifies ATT on the assumption that

  $E(y_{0it+1} \mid d_i = 1) = E(y_{0it} \mid d_i = 1)$
- However, even without the treatment, outcomes may have changed between t and t+1 because of macro factors or lifecycle effects.
- To control for these trends, we can include a control group who never receive the treatment but (are assumed to) experience the same trends.

Difference-in-difference estimator

The difference-in-difference (DID) estimator takes the difference between the change in outcomes for treated individuals and the change for untreated (control) individuals. DID is estimated as:

$$\left(\bar{y}_{t+1}^{T} - \bar{y}_{t}^{T}\right) - \left(\bar{y}_{t+1}^{C} - \bar{y}_{t}^{C}\right)$$

where $\bar{y}^{T}$ $(\bar{y}^{C})$ is the mean outcome for treated (control) individuals

A weakness of DID is that the common trend assumption may be violated:

- macro trends may affect the 2 groups differently
- may be time-varying factors affecting only one group, e.g. "Ashenfelter's dip": often trainees had a temp drop in earnings before they took up training course.

Regressions

Consider a regression model with a treatment dummy, time trend and interaction :

$$y_{it} = \alpha_0 + \gamma d_i + \theta w2_t + \rho d_i.w2_t + u_i + \varepsilon_{it},$$
$$t = 1, 2; i = 1 \ldots n$$

where $w2_t$ equals 1 if t=2 and zero otherwise.

It is easily shown that in this simple case (2 waves and no other controls) $\hat{\rho}$ is identical to DID and so identifies ATT. Can estimate as RE, FE (in which case d_i drops out) or by pooled OLS (adjust SEs).

Can add controls $\mathbf{x}_{it}$ to account for differing trends - though interpretation of $\hat{\rho}$ is less straightforward (unless treatment effect same for all, $\Delta_i = \Delta$).

Other estimators

- Other estimators of treatment effects match treatment and control individuals based on observed characteristics **x**. A popular estimator of this type is propensity score matching.
- Matching estimators can be less restrictive (don't assume linear functional form) and allow more flexible analysis of heterogeneous treatment effects.
- But they assume treatment is unrelated to potential outcomes conditional on **x**: selection on observables.
- Can also combine matching with DID.

Topic 6:

Count data models

Count data

- Quantities are often inherently discreet, or are measured discreetly. Frequencies are inherently discreet. Examples of count variables:
  - Number of visits to doctor
  - Number of organisations joined.
  - Number of arrests.
  - Number of patent applications.
- Counts cannot be negative, may be (are often?) zero and always take integer values.
- Modelling counts as continuous variables would not take account of this "lumpy" distribution (cf problems with LPM for binary variables).

Modelling count data (1)

Counts are typically modelled as a Poisson distribution. The probability of individual i experiencing y_{it} events in period t is:

$$\Pr(y_{it}) = \frac{\exp(-\lambda_{it})\lambda_{it}^{y_{it}}}{y_{it}!}$$

Where does this come from? Imagine a simple experiment that would produce a distribution of counts. We toss a coin n=10 times and count the number of heads (probability p of a head from a toss = 0.5).

This would produce a *binomial distribution*, with mean number of heads = np = 5.

Modelling count data (2)

- The Poisson distribution is the limiting form of the binomial distribution as the number of "trials" (tosses) goes to infinity, and p gets correspondingly smaller so as to keep constant the mean count np ($=\lambda$).
- The mean of the Poisson distribution is λ_{it}.
- The variance of the Poisson distribution is also λ_{it} (often rejected in practice!).
- Allow for observed and unobserved characteristics by specifying $\lambda_{it} = \exp(\mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i)$. Note the exponential form guarantees $\lambda_{it} > 0$.

Poisson regression

- The Poisson model is usually estimated by maximum likelihood (ML)
- The ML estimator is quite "robust": provided the conditional mean is correctly specified, the estimates are consistent even if the true distribution is not Poisson.
- The conditional mean is:

$$E(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) \equiv \lambda_{it} = \exp(\mathbf{z}_i \alpha + \mathbf{x}_{it} \beta + u_i)$$
$$= \exp(u_i) \,.\, \exp(\mathbf{z}_i \alpha + \mathbf{x}_{it} \beta)$$

Marginal effects (1)

- So the individual effect u_i affects the conditional mean multiplicatively. This turns out to be convenient.
- Since $E(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = \exp(u_i) \,.\, \exp(\mathbf{z}_i \alpha + \mathbf{x}_{it} \beta)$, if x_{jit} increases by 1 unit, holding all else constant, the ratio of the new to the old mean number of events is $\exp(\beta_j)$. In Stata, using the `irr` option, this is reported as an "incident rate ratio".
- Notice the IRR is independent of u_i.

Marginal effects (2)

- Alternatively, the marginal effect of x_{jit} on the expected count is:

$$\partial E(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) / \partial x_{jit} = \beta_j \exp(u_i).\exp(\mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta})$$
$$= \beta_j E(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i)$$

- Semi-elasticity: a 1 **unit** increase in x_{jit} has a 100 β_j % effect on the expected count, all else constant.
- Equivalently, β_j is the marginal effect on the log of the expected count.

Random effects Poisson model

- We still have to deal with u_i. In the RE model, we assume that the multiplicative individual effect is independent of $(\mathbf{z}_i, \mathbf{X}_i)$ and has a gamma distribution with a mean of one (analogous to mean zero in an additive model) and constant variance (=α in Stata).
- Stata also allows a normally distributed individual effect (but runs slower).

Fixed effects Poisson model

- The Poisson regression can also be estimated as a fixed effects model, allowing arbritary dependence of u_i on $(\mathbf{z}_i, \mathbf{X}_i)$.
- As for the conditional (FE) logit, the method is to condition on a sufficient statistic. The sufficient statistic is the sum for each individual of the observed counts over the panel ($= \sum_{t=1}^{T_i} y_{it}$)
- As usual in FE models, the effects (α) of time-invariant variables $\mathbf{z}_i$ cannot be identified.

Over- (and under-) dispersion

- A restrictive feature of the Poisson model is that the mean and variance of y_{it} are constrained to be the same.
- In practice, the variance is usually greater than the mean – overdispersion. One reason is unobserved heterogeneity (cf linear regression where individual effects increase the variance of the composite error term).
- The *negative binomial* distribution allows for overdispersion.
- But, with panel data techniques we already allow explicitly for unobserved heterogeneity.
  - RE incorporates overdispersion
  - FE is consistent in presence of either under- or overdispersion.