

## Day 3: Further topics selected from...

### (i) Binary response models

- Linear regression for binary variables
- Conditional (fixed-effects) logit and random effects logit/probit

### (ii) Dynamic panel data models

- Dynamic FE regression models: GMM methods
- Dynamic RE models for binary data

### (iii) Attrition and sample selection

### (iv) Policy evaluation in panels

## Day 3

### Topic (i) Binary response models

## 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*.

Here, we 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 the *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:

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

## Disadvantages of the LPM

The linear probability model requires:

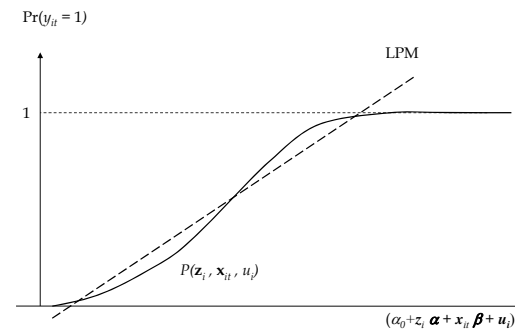
$$P_{it} \approx \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i$$

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

Moreover,  $\text{var}(y_{it} | \mathbf{z}_i, \mathbf{x}_{it}, u_i) = P_{it} [1 - P_{it}]$  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)

$$\begin{aligned} \Rightarrow \Pr(y_{it} = 1 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) &= \Pr(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it} > 0) \\ &= \Pr(-\varepsilon_{it} < [\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i]) \\ &= F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i) \end{aligned}$$

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

**Probit model:** assume  $\varepsilon_{it}$  has a normal distribution

$$F(\cdot) = \Phi(\cdot) \Rightarrow \text{df of the } N(0,1) \text{ distribution}$$

**Logit (logistic regression) model:** assume  $\varepsilon_{it}$  has a logistic distribution

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

### 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 =  $\alpha_0, \boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$

Predicted probability =  $P_{it}$

Odds ( $O_{it}$ ) =  $P_{it} / (1 - P_{it})$

For 2 people with different  $\mathbf{z}$  and  $\mathbf{x}$ -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  $(\mathbf{x}_0, u_0)$



08/11/2011 (33)



## Logistic regression and the odds ratio

In the logit model:

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

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

$$\text{Odds ratio} = [P_1 / (1 - P_1)] / [P_0 / (1 - P_0)]$$

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

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

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

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 on the probability  $\Pr(y=1)$ .



08/11/2011 (34)



## Misinterpretation of odds ratios

Check that you understand the error in the following quotation from a well-known textbook:

*"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."*

In fact, it isn't possible to calculate the relative risk or the marginal effect on the probability of poverty, from knowledge of the odds ratio alone.

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



08/11/2011 (35)



## Presentation of results

- We can report marginal effects evaluated at sample mean values of  $\mathbf{x}$  and  $\mathbf{z}$ , 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.
  - Almost no-one has a zero individual effect
- Present *average partial effects* (APE) which allow for the average effect of the unobserved individual effects. Evaluate at:
  - Mean  $\mathbf{x}$  and  $\mathbf{z}$ , or
  - Selected  $\mathbf{x}$  and  $\mathbf{z}$  to represent typical persons, or
  - Each sampled person's  $\mathbf{x}$  and  $\mathbf{z}$ , then average the results.
- These methods aren't possible with fixed-effects logit, as we don't estimate the (distribution of) individual effects or the coefficients of time-invariant variables.



08/11/2011 (36)



## Fixed effects models – some issues

- To deal with individual effects in linear FE models, we can:

- estimate individual effects  $u_i$  (LSDV).
- eliminate individual effects  $u_i$  by within-group transform

The two approaches are identical and give an unbiased estimate of  $\beta$

- But in non-linear FE models:
  - Can't remove the individual effects  $u_i$  by within-group transformation as in linear regression
  - With no short-cut method of calculating the estimate of  $\beta$  we'd have to use the dummy variable method and calculate estimates of all the  $u_i \Rightarrow$  the "incidental parameters problem"
  - All the estimated coefficients would be biased, even in large samples

## 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  $\beta$ .
- Then CML is consistent (loosely speaking, unbiased in a large sample of individuals) for  $\beta$ .
- But CML is 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)
- Details of conditional logit are given in Appendix 4

## Fixed effects (or conditional) logit

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

where  $F(\cdot)$  is the logistic form

Avoiding technicalities, the method works as follows:

- Use 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 who display no change in  $y$
- The changes in the covariates  $\mathbf{x}_{it}$  (i.e. variable differences like  $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$ ) are then used in a modified logit analysis to explain the changes in the observed sequence of outcomes  $y_{i1} \dots y_{iT}$ .
- NB differencing the covariates removes any variables constant over time (e.g. gender, birth year, etc.), so  $\alpha$  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 (unlike RE logit) is unaffected by any endogeneity which is confined to  $u_i$

## 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  $y_{it-1}$  as an explanatory variable)

In these circumstances, 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  $\mathbf{x}$  – 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  $\varepsilon_{it}$  are independent, the joint probability of observing  $(y_{i1}, \dots, y_{iT})$  conditional on  $u_i$  (and  $\mathbf{z}_i, \mathbf{x}_{it}$ ) is just the product of the conditional probabilities for each time period:

$$\begin{aligned} \Pr(y_{i1}, \dots, y_{iT} \mid \mathbf{z}_i, \mathbf{X}_i, u_i) \\ &= \Pr(y_{i1} \mid \mathbf{z}_i, \mathbf{x}_{i1}, u_i) \times \dots \times \Pr(y_{iT} \mid \mathbf{z}_i, \mathbf{x}_{iT}, u_i) \\ &= F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{i1} \boldsymbol{\beta} + u_i) \times \dots \times F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{iT} \boldsymbol{\beta} + u_i) \end{aligned}$$



08/11/2011 (21)



## Random effects logit/probit

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

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

$\Pr(y_{i1}, \dots, y_{iT} \mid \mathbf{z}_i, \mathbf{X}_i) = E[\Pr(y_{i1}, \dots, y_{iT} \mid \mathbf{z}_i, \mathbf{X}_i, u_i)]$   
where "E[.]" refers to the expectation or mean with respect to the  $N(0, \sigma_u^2)$  distribution of  $u_i$ .

This unconditional probability  $\Pr(y_{i1}, \dots, y_{iT} \mid \mathbf{z}_i, \mathbf{X}_i)$  is the likelihood for individual  $i$ . This process is repeated 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 can be quite long.

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



08/11/2011 (22)



## 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 Conditional Logit).
- $\hat{\boldsymbol{\beta}}_{CL}$  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:

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

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



08/11/2011 (23)



## Individual effects correlated with regressors

- The RE probit/logit assumes that  $(\mathbf{z}_i, \mathbf{x}_{it})$  and  $u_i$  are independent.
- Ideal solution is to start from a theory to account for the  $\mathbf{x}_{it} - u_i$  endogeneity and estimate an appropriate simultaneous model
- A crude theory-free alternative is to allow  $u_i$  to be correlated with elements of  $\mathbf{x}_{it}$  observed in the sample:
  - General formulation due to Chamberlain models the mean of  $u_i$  as a function of all values of  $\mathbf{x}_{it}$  from all time periods.
  - Simplified version (based on the Mundlak model) is to model  $u_i$  as a function of individual means,  $\bar{\mathbf{x}}_i$ :
 
$$u_i = \mu + \bar{\mathbf{x}}_i \boldsymbol{\delta} + \eta_i, \quad \text{where } \eta_i \mid \bar{\mathbf{x}}_i \sim N(0, \sigma_\eta^2)$$
  - NB: this cannot be a structurally stable model



08/11/2011 (24)



### Unobserved heterogeneity or state dependence?

- As seen in the HILDA data, 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 plausible 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.

## Day 3

### Topic (ii) Dynamic models

## Dynamic models

Why model dynamics?

- Current outcomes might depend on past values of determinants  $\Rightarrow$  include lagged  $\mathbf{x}$  (distributed lag model). Use the techniques already discussed.
- Adjustment might be partial: this year's outcome  $y$  depends not only on  $\mathbf{x}$ , but also on last year's outcome  $\Rightarrow$  include lagged  $y$ . We focus on this case.
- Note (as we'll see) this amounts to including an infinite (or back to start of process) number of lagged  $\mathbf{x}$ , so it can accommodate longer-range dependence.

### 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):

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

or static model with AR(1) errors:

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

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

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

NB: model (1) implies gradual adjustment to change in  $\mathbf{x}$ ; 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)):

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

where:

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

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 on

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

## 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  $y$ ).

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

By substitution:

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

and so on... eventually we arrive at the point of origin (e.g. the individual's birth),  $t=0$

## Properties of the within-group estimator (2)

Distributed lag form of (1):

$$\begin{aligned} y_{it} &= \sum_{s=0}^{t-1} \gamma^s (\mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{i,t-s} \boldsymbol{\beta} + u_i + \varepsilon_{i,t-s}) + \gamma^t y_{i0} \\ &= \frac{1-\gamma^t}{1-\gamma} (\mathbf{z}_i \boldsymbol{\alpha} + u_i) + \sum_s \gamma^s \mathbf{x}_{i,t-s} \boldsymbol{\beta} + [\varepsilon_{it} + \gamma \varepsilon_{i,t-1} + \dots + \gamma^{t-1} \varepsilon_{i1}] + \gamma^t y_{i0} \end{aligned}$$

$\Rightarrow y_{it-1}$  is a function of  $\varepsilon_{i,t-1} \dots \varepsilon_{i1}$

$\Rightarrow \bar{y}_i^* = \sum_{t=0}^{T_i-1} y_{it} / T_i$  is a function of  $\varepsilon_{i,T-1} \dots \varepsilon_{i1}$  and  $y_{i0}$

$\Rightarrow y_{i,t-1} - \bar{y}_i^*$  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  $\gamma$  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:  

$$y_{it} = \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it-1} + u_i + \varepsilon_{it}$$
- 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 pooled OLS is biased and inconsistent as  $n \rightarrow \infty$

## 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. All are available in Stata.
  - Anderson-Hsiao
  - Arellano-Bond
  - Blundell-Bond
- An alternative approach, due to Kiviet, is to adjust the FE regression to eliminate the bias – also available in Stata

## A simple IV estimator

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

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

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

- ⇒ Find a set of instruments correlated with  $\Delta y_{it-1}$  but uncorrelated with  $\varepsilon_{it-1}$
- ⇒ All lagged  $\mathbf{x}_{it}$  and  $y_{it-2} \dots y_{i0}$  are valid instruments if  $\{\varepsilon_{it}\}$  is serially independent
- ⇒ Simplest IV estimator (Anderson Hsiao) estimates (1), using instruments  $(\mathbf{x}_{it}, \mathbf{x}_{it-1}, \mathbf{x}_{it-2}, y_{it-2})$ .
- ⇒ We can only use observations  $t = 2 \dots T_i$ . Each extra lag used as an instrument loses us  $n$  observations.
- ⇒ 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).

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

⇒ 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:

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

This is a system of  $T_i - 1$  linear equations with cross-correlated errors (since  $\Delta \varepsilon_{it}$  is correlated with  $\Delta \varepsilon_{it-1}$  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  $(\mathbf{x}_{i0} \dots \mathbf{x}_{iT}, y_{i0})$ ;  
the equation for  $t = T_i$  has  $(\mathbf{x}_{i0} \dots \mathbf{x}_{iT}, y_{i0} \dots y_{iT-2})$ .

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  $\hat{m}(y; \mu) = \frac{1}{n} \sum_{i=1}^n (y_i - \mu) = 0$
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  $\mathbf{V}$  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,  $\hat{\beta}_{IV} = (\mathbf{X}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}$   
 is the following M-estimator:

$$\begin{aligned}\hat{\beta}_{IV} &= \arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)' \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1} \mathbf{Q}'(\mathbf{y} - \mathbf{X}\beta) \\ &= \arg \min_{\beta} \hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}, \beta)' \mathbf{V}^{-1} \hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}, \beta)\end{aligned}$$

where  $\hat{\mathbf{m}}$  is the "sample analogue",  $n^{-1}\mathbf{Q}'(\mathbf{y}-\mathbf{X}\beta)$ , 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  $\sigma_{\varepsilon}^2$  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:

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

Full set of moment conditions:

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

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

.

.

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

$\hat{\mathbf{m}}$  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  $E\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i'$

## Assessment of the GMM method

- Very many 'internal' instruments/moment conditions
- Even more conditions can be added (e.g. for  $\mathbf{z}_i$  and to impose the homoskedasticity assumption on  $\varepsilon_{it}$ )
- **But** GMM often works badly in finite samples with many moment conditions - a leading example of the "weak instruments problem" (see Roodman, *Oxf. Bull. Econ. & Stat.*, 2009)
- In practice, this means results can often be very unstable - small changes in specification, choice of instruments, etc. can make large differences to the results

## 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

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

has an asymptotic chi-square distribution with  $r - k_x$  degrees of freedom (Hansen-Sargan test).

## Specification testing

### (2) Testing for residual serial correlation

If the  $\varepsilon_{it}$  are serially independent, then

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

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

Thus, the first order serial correlation coefficient is

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

But  $E[\Delta \varepsilon_{it} \Delta \varepsilon_{it-2}] = 0$ , and so the second order serial correlation coefficient  $r_2 = 0$ .

$\Rightarrow$  test for second order serial correlation.

There is a 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  $\Delta y_{i2} \dots \Delta y_{iT}$ .

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:

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

$\Rightarrow$  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  $\Delta y_{it-1}$ , 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

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

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

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

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

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

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

## Dynamic random effects probit

- We focus on the RE probit model with a simple dynamic specification (one lag of the dependent variable).
- The latent regression is now:
 
$$y_{it}^* = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it-1} + u_i + \varepsilon_{it}$$

$$y_{it} = 1 \text{ if } y_{it}^* > 0 \text{ and } y_{it} = 0 \text{ otherwise}$$
- True state dependence is measured by  $\gamma$ , and persistent unobserved heterogeneity is captured by  $u_i$
- Assume (as previously) that  $\varepsilon_{it}$  is serially uncorrelated [otherwise we are neglecting some persistent unobserved heterogeneity].

## The random effects likelihood function

Construct a likelihood by sequential conditioning. Define in turn:

$$P_{i0} = \Pr(y_{i0} \mid \mathbf{z}_i, \mathbf{X}_i, u_i)$$

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

⋮

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

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

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

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

Given a particular value for  $u_i$ , the likelihood function for individual  $i$  is:

$$L_i(u_i) = P_{i0}(u_i) \times P_{i1}(u_i) \times \dots \times P_{iT}(u_i)$$

## 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 (i.e. unrelated to the individual effect) then effectively  $P_{i0}$  can be dropped from the likelihood
  - Possible efficiency loss since useful information about the starting point may be neglected.
- But  $y_{i0}$  is probably not exogenous:
  - It may not be the true starting point of the “process”, just the start of our sample
  - In any case,  $y_{i0}$  may not be 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:

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

and the latent regression is now:

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

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.

## Day 3

### Topic (iii) Attrition and sample selection

### 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've 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 missingness:
  - Missing completely at random (MCAR)
  - Missing at random (MAR): conditional on covariates ( $X_i, z_i$ ), response is random. Systematic differences in response are explained by observable characteristics.
  - Non-ignorable non-response: systematic differences in response remain after controlling for ( $X_i, z_i$ ).

### Implications of incompleteness

- Implications depend on type of analysis.
- Descriptive (i.e. unconditional) statistics like sample means and proportions 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 otherwise.
  - Example: if poor households are less likely to participate in surveys, a regression of consumption on income will be unbiased, provided income is exogenous with respect to consumption behaviour.

## 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;
  - Calibration to align sample with (e.g.) census data
- Use weights for descriptive stats (if want to make inferences about the population).
- Weighting is more problematic in regression analysis:
  - General purpose weighting 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 only inflates SEs.
  - If weights have a large effect, it may indicate model misspecification rather than non-response problems
  - In practice, Stata does not accept weights for linear FE and RE (GLS) analysis.

## Non-random selection in panels

- Panel regression model:

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

- Define a response indicator:

$$r_{it} = \begin{cases} 1 & \text{if } (y_{it}, \mathbf{z}_i, \mathbf{x}_{it}) \text{ is observed at wave } t \\ 0 & \text{otherwise.} \end{cases}$$

- If data are MCAR or MAR, then  $r_{it}$  is independent of  $u_i$  and  $\varepsilon_{it}$ , conditional on the covariates  $\mathbf{z}_i, \mathbf{X}_i$
- If non-response is non-ignorable then  $r_{it}$  is not independent of  $u_i$  and  $\varepsilon_{it}$ . Also called *selection on unobservables*.

## Consequences for RE estimates

- Implications of missing observations for linear RE and FE estimates.
- RE is unbiased if:  
$$E(u_i + \varepsilon_{it} | \mathbf{X}_i, \mathbf{z}_i, \mathbf{r}_i) = E(u_i + \varepsilon_{it} | \mathbf{X}_i, \mathbf{z}_i) = 0$$
where  $\mathbf{r}_i = (r_{i1}, \dots, r_{iT})$ , 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

- FE is robust to certain forms of non-random selection into the panel.
- FE is unbiased if:  
$$E(\varepsilon_{it} | \mathbf{X}_i, \mathbf{r}_i) = E(\varepsilon_{it} | \mathbf{X}_i) = 0$$
This says that the transitory error term is unrelated to selection, conditioning on time-varying observable characteristics. But  $\mathbf{r}_i$  can be related to  $u_i$ .
- As long as selection into the panel is related only to time-invariant factors, FE remains consistent.
- But, if  $r_{it}$  is related to the time-varying residual  $\varepsilon_{it}$ , FE is biased (and might even be more biased than RE)

## Testing for non-random selection in panels

- Simple indicative tests for non-random selection:
  - check whether  $r_i$  helps explain the outcome  $y_{it}$  after controlling for other characteristics
  - compare results from the unbalanced panel with the balanced sub-panel.
- In the first type of test, functions of  $r_i$  can be added to the equation and their significance tested [note the current  $r_{it}$  can't be added – why not?], e.g. :
  - lagged response indicator,  $r_{it-1}$
  - indicator for presence in all waves,  $c_i = \min(r_{i1} \dots r_{iT})$
  - number of waves present,  $T_i = \sum r_{it}$The last two can only be used with RE since  $c_i$  and  $T_i$  are time-invariant.

## Hausman test

Another test compares RE or FE estimates from the unbalanced panel and its balanced sub-panel. If selection is random, they should be close. Otherwise, there may be a statistically significant difference between the two.

For example, test the RE estimator using the statistic:

$$\left( \hat{\beta}_{RE,B} - \hat{\beta}_{RE,U} \right)' \left[ \text{var}(\hat{\beta}_{RE,B}) - \text{var}(\hat{\beta}_{RE,U}) \right]^{-1} \left( \hat{\beta}_{RE,B} - \hat{\beta}_{RE,U} \right) \\ \sim \chi^2(k) \text{ under } H_0 : \text{no selection bias}$$

May not have good power, since both estimates may be affected similarly by attrition.

If tests suggest attrition bias, the situation is difficult: methods to correct for “endogenous” attrition are complicated and depend on a correct understanding of the attrition process.

## Day 3

### Topic (iv) Policy evaluation in panels

## 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  $\Delta_i = y_{1i} - y_{0i}$ . Note that  $\Delta_i$  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):  $E(y_{1i} | d_i = 1) - E(y_{0i} | d_i = 0)$   
 $= E(y_{1i} | d_i = 1) - E(y_{0i} | d_i = 1) + E(y_{0i} | d_i = 1) - E(y_{0i} | d_i = 0)$   
 $= E(y_{1i} - y_{0i} | d_i = 1) + E(y_{0i} | d_i = 1) - E(y_{0i} | d_i = 0)$   
 $= ATT + E(y_{0i} | d_i = 1) - E(y_{0i} | d_i = 0)$   
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  $E(y_{0i} | d_i = 1) - E(y_{0i} | d_i = 0)$ , 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:

$E(y_{1it+1} | d_i = 1) - E(y_{0it} | d_i = 1)$   
by  $\bar{y}_{t+1}^T - \bar{y}_t^T$ , 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
 
$$E(y_{0it+1} | d_i = 1) = E(y_{0it} | d_i = 1)$$
- 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:

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

where  $\bar{y}^T$  and  $\bar{y}^C$  are mean outcomes 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 :

$$y_{it} = \alpha_0 + \gamma d_i + \theta w_{2t} + \rho d_i \cdot w_{2t} + u_i + \varepsilon_{it},$$

$$t = 1, 2; i = 1 \dots n$$

where  $w_{2t}$  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).

Can add controls  $x_{it}$  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 is propensity score matching (Rosenbaum & Rubin, *Biometrika* 1983).
- 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.

## Appendix 4

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

## 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:

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

## Marginal effects (2)

- Marginal effect is coefficient multiplied by the density function (normal for probit, logistic for logit), evaluated at the base values of  $\mathbf{x}$ .
- 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  $f(\cdot)$  cancels out). So the ratio of MEs due to  $x_j$  and  $x_k$  is  $\beta_j / \beta_k$ . 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  $u_i$ :

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

The statistic  $\sum_t y_{it}$  is a sufficient statistic for  $u_i$ :  $\Pr(y_i | \sum_t y_{it})$  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:

$$\begin{aligned} \Pr(\sum_t y_{it} = 1) &= \Pr(y_{i1} = 1, y_{i2} = 0) + \Pr(y_{i1} = 0, y_{i2} = 1) \\ &= P_{i1}(1 - P_{i2}) + (1 - P_{i1})P_{i2} \\ &= \frac{e^{x_{i1}\beta + u_i} + e^{x_{i2}\beta + u_i}}{(1 + e^{x_{i1}\beta + u_i})(1 + e^{x_{i2}\beta + u_i})} \end{aligned}$$

Conditional probability:

$$\begin{aligned} \Pr(y_{i1} = 1, y_{i2} = 0 \mid y_{i1} + y_{i2} = 1) &= \frac{\Pr(y_{i1} = 1, y_{i2} = 0)}{\Pr(y_{i1} + y_{i2} = 1)} \\ &= \frac{e^{x_{i1}\beta + u_i}}{e^{x_{i1}\beta + u_i} + e^{x_{i2}\beta + u_i}} = \frac{e^{x_{i1}\beta}}{e^{x_{i1}\beta} + e^{x_{i2}\beta}} = \frac{e^{(x_{i1} - x_{i2})\beta}}{1 + e^{(x_{i1} - x_{i2})\beta}} \end{aligned}$$

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

## Conditional logit (continued)

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

$$L(\beta) = \sum_{i: \sum_t y_{it} = 1} (d_i (x_{i1} - x_{i2})\beta - \ln(1 + e^{(x_{i1} - x_{i2})\beta}))$$

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

Note that, if  $x_{it}$  contains time-invariant covariates (i.e.  $z_i$ ), these disappear from  $(x_{i1} - x_{i2}) \Rightarrow \alpha$  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  $P_{it}(u_i) = \Pr(y_{it} \mid z_i, x_{it}, u_i)$ , where

$$\Pr(y_{it} \mid z_i, x_{it}, u_i) = \begin{cases} F(\alpha_0 + z_i\alpha + x_{it}\beta + u_i) & \text{if } y_{it} = 1 \\ 1 - F(\alpha_0 + z_i\alpha + x_{it}\beta + u_i) & \text{if } y_{it} = 0 \end{cases}$$

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

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

which tells us, for given values of  $\alpha, \beta, \sigma_u^2$  and  $\sigma_\varepsilon^2$ , 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  $L = \prod L_i$

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