Panel Data Models

1. Introduction

- Pure cross-section: Sample of individuals/firms/industries/households.
- Pure time-series: Sample over time.
- Panel follows the same sample of individuals/ firms/ industries/ households etc. over time.
- i.e. have multiple observations per cross-section unit.

i.e. Has two dimensions: cross-section and a time-series.

Other similar setups:

- Information on twins or siblings in families
- Employees in different firms.

<u>Benefits</u>

- Can model more complicated individual behaviour. Can control for individual heterogeneity unlike pure TS. No aggregation bias.
- Can study dynamics (given sufficient length over time).
- The <u>sequencing</u> of events enables us to study causal effects.

- Unobservable individual specific or time specific effects can be allowed for.

  Consider $w_i = \beta_1 + \beta_2 S_i + \beta_3 A_i + \text{error}$

  Bias in $\hat{\beta}_2$ for $\beta_2 = \beta_3$ [coeff. reg of A_i on S_i] = [+ve] [+ve] = +ve.

- More informative data, more variability, more df, less multicoll. problems.

Limitations

- Very expensive to collect.

- Problems of attrition in long panels. Final sample may not be representative. Beware of endogenous attrition!
- Recall problems with retrospective panels.
- Measurement error problems - interpretation of questions; recall; (in some cases the bias due to measurement errors might be more compared to pure c-s analysis).

Types of longitudinal data

- pseudo panel
- retrospective survey

- prospective survey (limited/unlimited duration)
- admin data

Other considerations

Incomplete panels: why is it incomplete? Issues of non-randomly missing data, attrition issues, selection bias….

Unbalanced panels: same as above….

Rotating panels: generally no problems….

GENERAL MODEL

$y_{it} = c_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + u_{it}$ $\quad i=1,..,N; \quad t=1,..,T$

$$y_{it} = c_i + \mathbf{x}_{it}\,\boldsymbol{\beta} + u_{it} \qquad (1)$$

[Wooldridge notation!]

- $\mathbf{x}_{it}$ (1 x k) will include time varying as well as time invariant variables.
- c_i unobserved (heterogeneity, indiv effect, etc. for ability, motivation,.)

  [no mention of random vs fixed effects yet! Just unobserved effects]

- can have a δ_t (use dummies when T is small)

- Assume we have a balanced panel. If it is unbalanced, we will need to make sure that it is not because of some kind of selection!

- Assume random sampling (**independently drawn**) in the cross-section dimension (what if c is geographical regions?)

- Typically, we will deal with large N and small T panels. Asymptotics handled via fixed T and as $N \rightarrow \infty$. Time series properties not relevant (can have non-stationarity)

Example: y_{it} - log earnings of individual i in time t. (i) Number of unemployment spells in each time period; (ii) Education, sex, ethnicity; (iii) Unemployment rate at the aggregate level.

GENERAL NOTATION

$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}$ [$\mathbf{x}_{it}$ is 1xK vector; β is Kx1 vector]

$\mathbf{y}_i = \mathbf{X}_\mathbf{i}\boldsymbol{\beta} + c_i + u_\mathbf{i}$ **[**$\mathbf{y}_i$ is T x 1; $\mathbf{X}_\mathbf{i}$ is T x K]

So $\mathbf{X}_i'\mathbf{X}_i$ will be K x K.

[l.c bold is a vector; u.c bold is a matrix]

ISSUES OF EXOGENEITY OF THE Xs

3 components: $\mathbf{x}_{\mathbf{it}}$, c_i, u_{it}

timing is important here.

All conditional on c_i

[can make assumptions without conditioning on c but: does it make sense?]

<u>Strict exogeneity</u>: [very strong assumption]

$E(u_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = 0 \qquad \forall\, t$ [note future values]

$E(y_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it} \mid \mathbf{x}_{it}, c_i) = \mathbf{x}_{it}\beta + c_i$

Note: once you control for $\mathbf{x}_{it}$ **and** c_i: $\mathbf{x}_{is}$ **(s** $\neq$ t) has no partial effect on y_{it}.

We say that $\{\mathbf{x}_{it}$**:** $t=1,..,T\}$ are strictly exog conditional on c_i.

Assumption will fail in **LDV** models! [more later!]

Weak exogeneity or sequential exogeneity (predetermined regressors)

$E(u_{it} \mid \mathbf{x_{i1}}, \mathbf{x_{i2}}, \ldots, \mathbf{x_{it}}, c_i) = 0 \qquad \forall\, t \qquad$ [no future values of **x**]

$$\boxed{E(y_{it} \mid \mathbf{x_{i1}}, \mathbf{x_{i2}}, \ldots, \mathbf{x_{it}}, c_i) = E(y_{it} \mid \mathbf{x_{it}}, c_i) = \mathbf{x_{it}}\beta + c_i}$$

Ok in LDV models but will fail in models with endogenous regressors.

Contemporaneous exogeneity:

$E(u_{it} \mid \mathbf{x_{it}}, c_i) = 0 \qquad \forall\, t \qquad [\mathbf{x_{it}} = (x_{i1}, x_{i2}, \ldots, x_{ik})]$

$$\boxed{E(y_{it} \mid \mathbf{x_{it}}, c_i) = \mathbf{x_{it}}\beta + c_i}$$

<u>Unconditionally on c</u>

question is what is $E(c_i \mid \mathbf{x_{i1}, x_{i2}, \ldots, x_{iT}})$

This is in general $\neq E(c_i)$ [this is what we have to bear in mind]

NOTES

The above zero conditional expectations imply the following:

<u>Strict exogeneity</u>: $E(\mathbf{x_{it}}'u_{is})=0, \quad \forall\ t,s$ [Uncorr across all t and s.]

<u>Weak exogeneity:</u> $E(\mathbf{x_{is}}'u_{it})=0 \quad s=1,\ldots,t$ [Uncorr with past **x.**]

<u>Contemporaneous exogeneity</u>: $E(\mathbf{x_{it}}'u_{it})=0 \quad \forall\ t$ [Contemp uncorr]

IMPORTANT QUESTIONS TO ASK

What is the most reasonable assumption to make? This has implications for the estimation technique one uses.

1. In general $E(c_i \mid \mathbf{x_{i1}, x_{i2}, \ldots, x_{iT}}) \neq E(c_i)$; **[LDV.]**
2. The most restrictive strict exogeneity (conditional on c) assumption will **not** hold if there is correlation between u_{it} and perhaps one of the future values of one of the $\mathbf{x_i}$**s. [LDV model or endogenous regressor (programme participation).]**