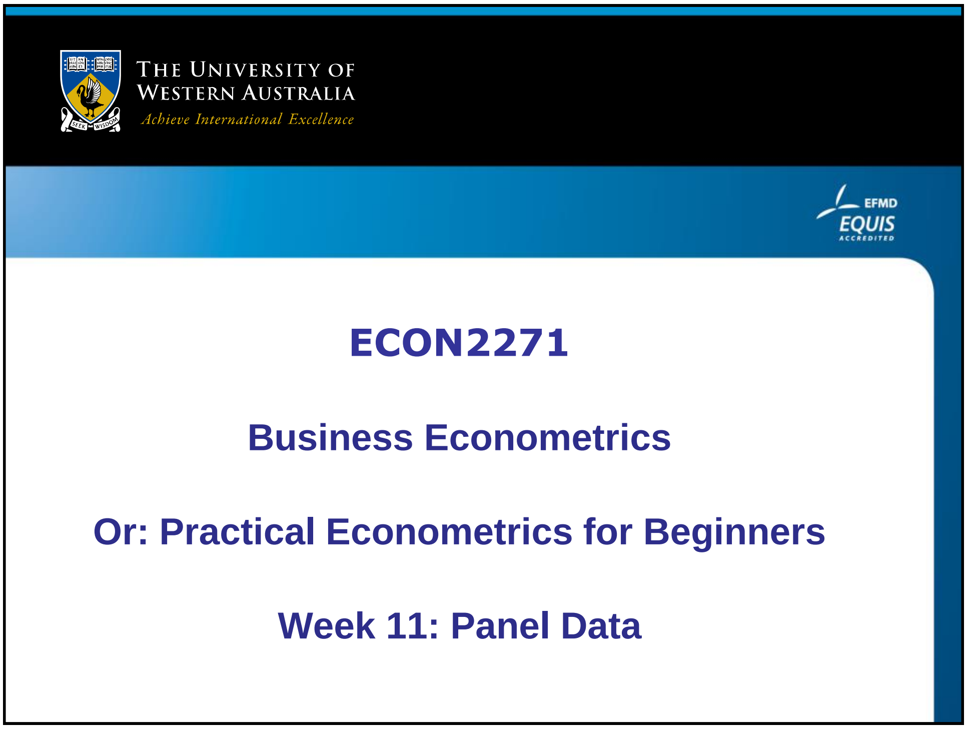


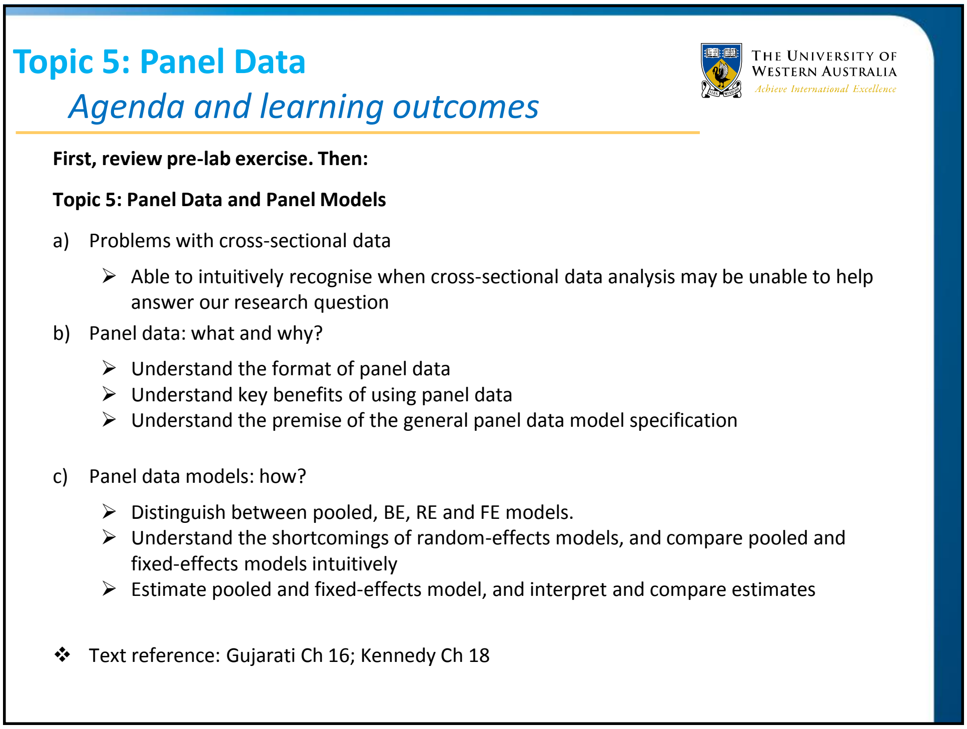
THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

ECON2271

Business Econometrics

Or: Practical Econometrics for Beginners

Week 11: Panel Data



Topic 5: Panel Data
Agenda and learning outcomes

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

First, review pre-lab exercise. Then:

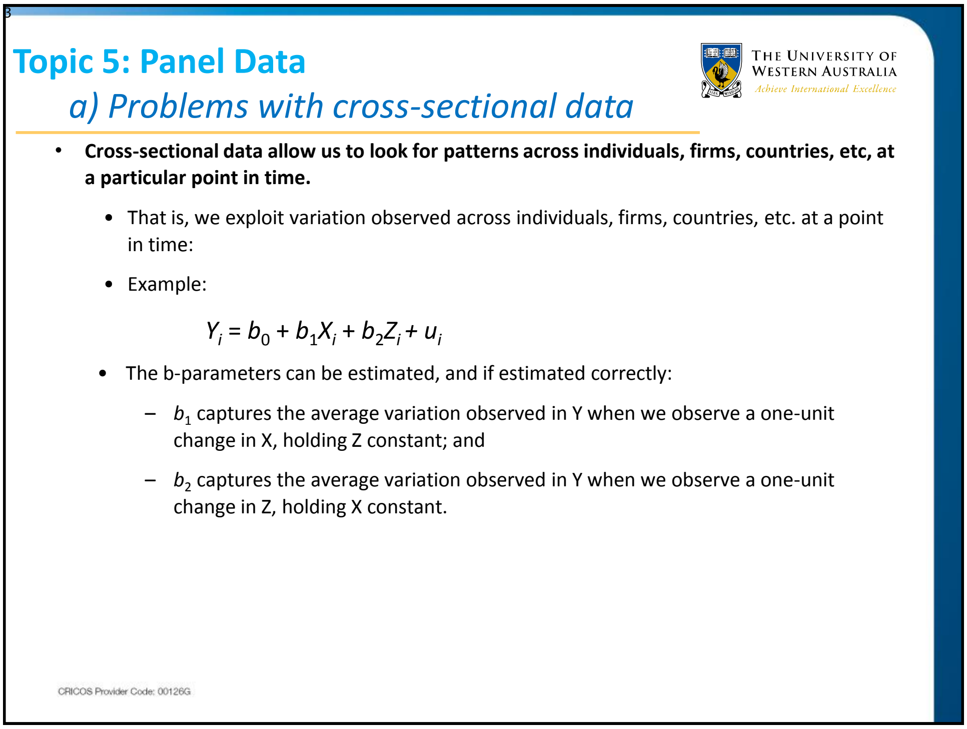
Topic 5: Panel Data and Panel Models

a) Problems with cross-sectional data
   - ➢ Able to intuitively recognise when cross-sectional data analysis may be unable to help answer our research question

b) Panel data: what and why?
   - ➢ Understand the format of panel data
   - ➢ Understand key benefits of using panel data
   - ➢ Understand the premise of the general panel data model specification

c) Panel data models: how?
   - ➢ Distinguish between pooled, BE, RE and FE models.
   - ➢ Understand the shortcomings of random-effects models, and compare pooled and fixed-effects models intuitively
   - ➢ Estimate pooled and fixed-effects model, and interpret and compare estimates

❖ Text reference: Gujarati Ch 16; Kennedy Ch 18



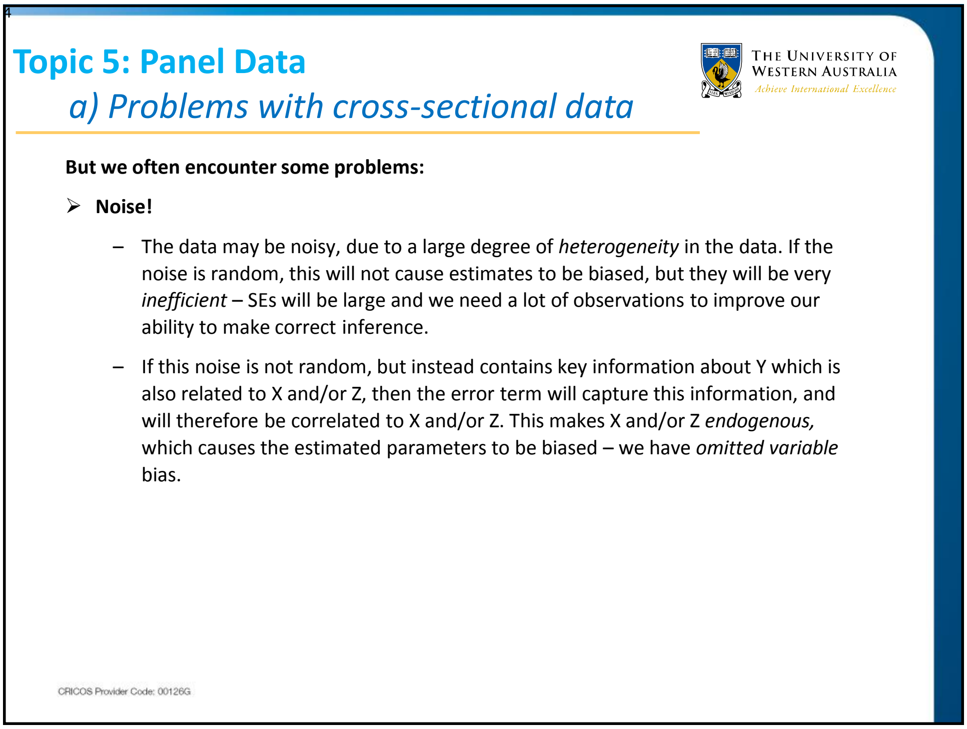
Topic 5: Panel Data

a) Problems with cross-sectional data

- **Cross-sectional data allow us to look for patterns across individuals, firms, countries, etc, at a particular point in time.**
  - That is, we exploit variation observed across individuals, firms, countries, etc. at a point in time:
  - Example:

$$Y_i = b_0 + b_1X_i + b_2Z_i + u_i$$

  - The b-parameters can be estimated, and if estimated correctly:
    - b_1 captures the average variation observed in Y when we observe a one-unit change in X, holding Z constant; and
    - b_2 captures the average variation observed in Y when we observe a one-unit change in Z, holding X constant.



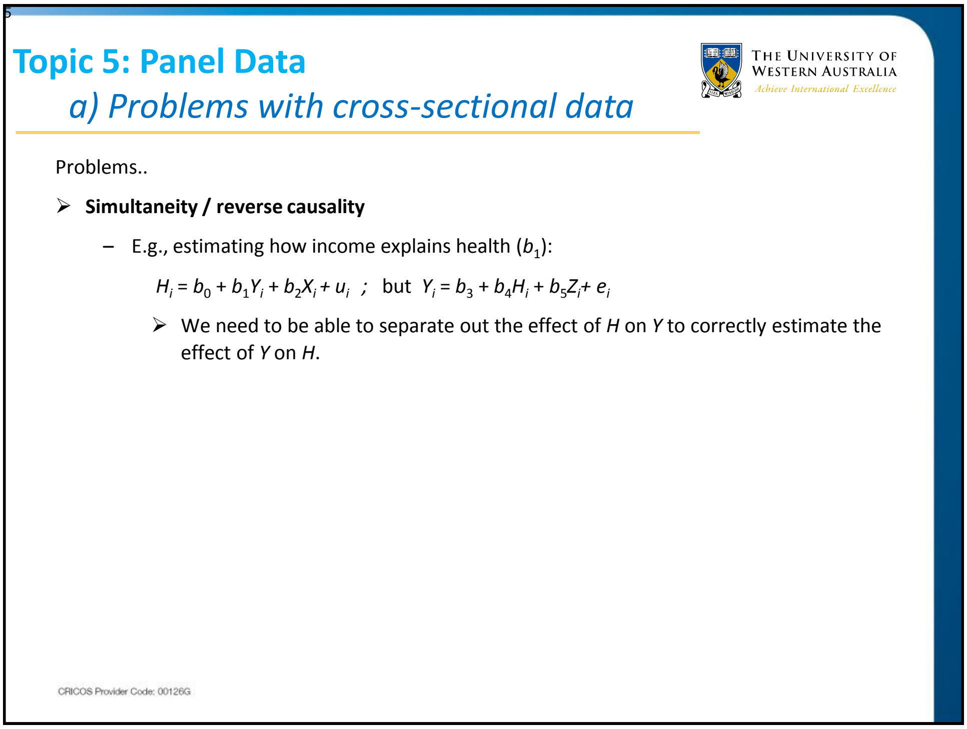
Topic 5: Panel Data

a) Problems with cross-sectional data

But we often encounter some problems:

➢ **Noise!**

- The data may be noisy, due to a large degree of *heterogeneity* in the data. If the noise is random, this will not cause estimates to be biased, but they will be very *inefficient* – SEs will be large and we need a lot of observations to improve our ability to make correct inference.
- If this noise is not random, but instead contains key information about Y which is also related to X and/or Z, then the error term will capture this information, and will therefore be correlated to X and/or Z. This makes X and/or Z *endogenous,* which causes the estimated parameters to be biased – we have *omitted variable* bias.



Topic 5: Panel Data

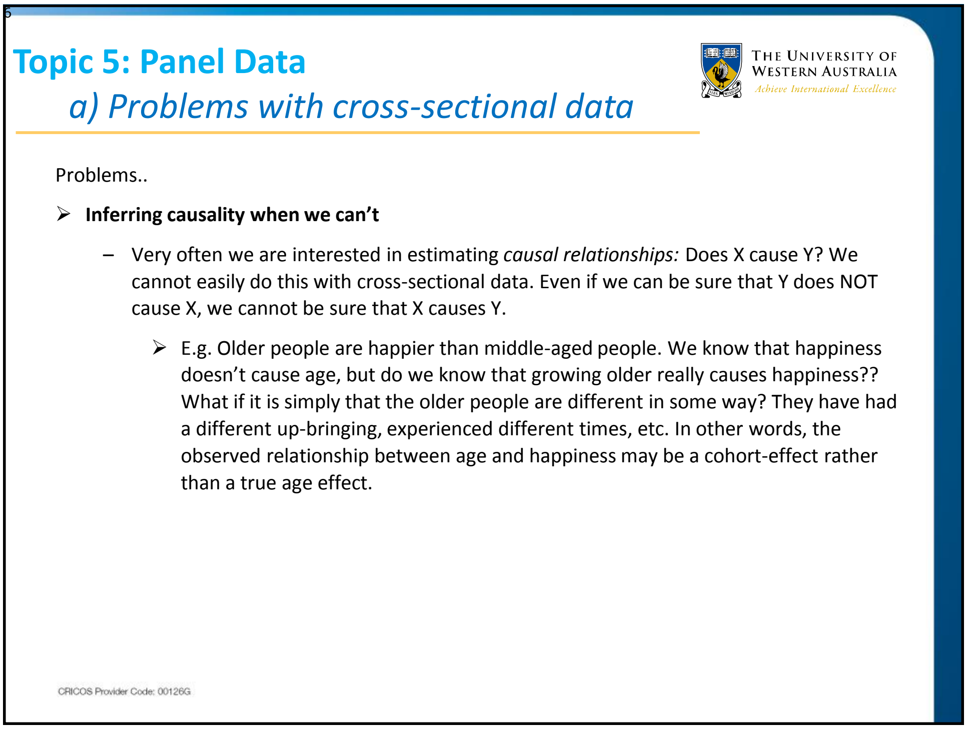
a) Problems with cross-sectional data

Problems..

- **Simultaneity / reverse causality**
  - E.g., estimating how income explains health (b_1):

    $H_i = b_0 + b_1Y_i + b_2X_i + u_i$; but $Y_i = b_3 + b_4H_i + b_5Z_i + e_i$

    - We need to be able to separate out the effect of *H* on *Y* to correctly estimate the effect of *Y* on *H*.



Topic 5: Panel Data

a) Problems with cross-sectional data

Problems..

- **Inferring causality when we can’t**
  - Very often we are interested in estimating *causal relationships:* Does X cause Y? We cannot easily do this with cross-sectional data. Even if we can be sure that Y does NOT cause X, we cannot be sure that X causes Y.
    - E.g. Older people are happier than middle-aged people. We know that happiness doesn’t cause age, but do we know that growing older really causes happiness?? What if it is simply that the older people are different in some way? They have had a different up-bringing, experienced different times, etc. In other words, the observed relationship between age and happiness may be a cohort-effect rather than a true age effect.

7

Topic 5: Panel Data

a) Problems with cross-sectional data

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Do we have solutions?**
  - ➢ **Too much noise: Obtain more and better variables?**
    - This can improve our ability to capture as much of the heterogeneity across individuals (or firms or countries etc etc) as possible.
    - For example, such heterogeneity is often referred to as *personal characteristics,* but many large surveys now collects information on personality. These data do not capture all this heterogeneity, but it can capture a lot if ot.
    - For example, Boyce (2010; J of Ec. Psych.) finds that a significant portion of these “individual fixed effects” (= heterogeneity) is accounted for by the Big 5 personality inventory (extraversion, agreeableness, conscientiousness, emotional stability and openness to experience).
    - Hence, we can potentially fix a lot problems associated with heterogeneity by being able to include such information.

CRICOS Provider Code: 00126G

8

Topic 5: Panel Data

a) Problems with cross-sectional data

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Solutions?

➢ **Simultaneity / reverse causality: Can we separate out what we need?**

– E.g., estimating how income explains health (b_1):

$H_i = b_0 + b_1 Y_i + b_2 \boldsymbol{X}_i + e_i$; but $Y_i = b_3 + b_4 H_i + b_5 \boldsymbol{Z}_i + u_i$

- ➢ We need to be able to separate out the effect of *H* on *Y* to correctly estimate the effect of *Y* on *H*.
- ➢ We may be able to use 2-Stage Least Squares (2SLS), which involves constructing an instrumental variable (IV) within the model:

1. Stage 1 – construct an IV for Y by regressing Y on all exogenous variables in the model:

   $Y_i = \hat{b}_6 + \hat{b}_7 X_i + \hat{b}_8 Z_i + v_i \Rightarrow \hat{Y}_{\mathrm{i}}$ = the IV for Y = the part of Y explained by *X*’s and *Z*’s (but not by H.. – hopefully, that part of the variation, which we don’t want to include, is contained in the error term v_i)

2. Stage 2 – estimate $H_i = b_9 + b_{10} \hat{Y}_{\mathrm{i}} + b_{11} \boldsymbol{X}_i + e_i$

- ➢ If this all goes to plan, the b_{10} estimate is now an unbiased (“pure”) estimate of the “effect of *Y* on *H*.
- ➢ BUT: this depends on how well the first stage (reduced form) regression is estimated
- ➢ Text ref: Gujarati Ch. 20, section 20.4

Topic 5: Panel Data

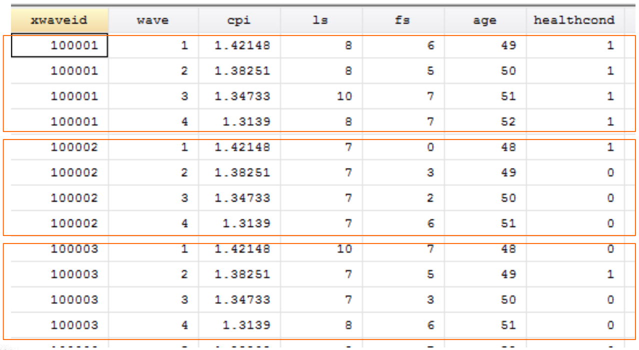
a) Problems with cross-sectional data

- **In summary:**
  - We can (and should!) do as much as we can to solve the problems inherent in using cross-sectional data whenever we are interested in anything more than comparing differences in Y with difference in X across individuals, firms, countries, etc.
  - This is particularly true when making inferences about causal relationships, but also when there is a possibility that some important information is omitted from the model (which is often the case).
  - Many of these problems can be addressed if we can observe variation *within* individuals as well as variation *across* individuals.
  - That is, we need to add another dimension to the data: time as well as space!

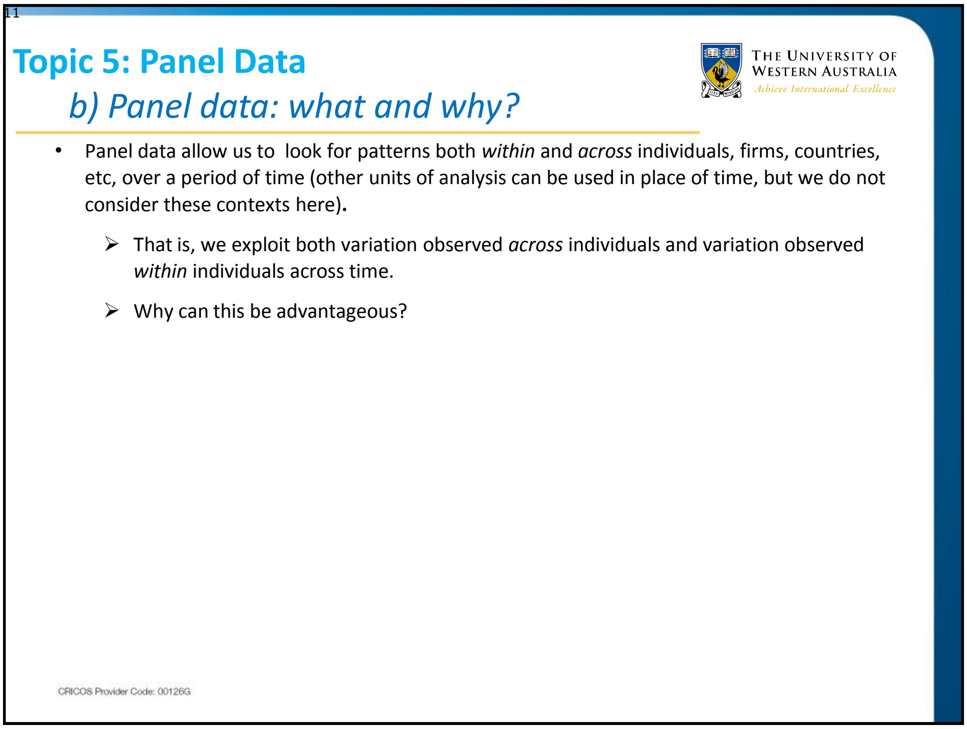
Topic 5: Panel Data

b) Panel data: what and why?

- Cross-sectional data comprises observations on a set of variables across a sample of units (individuals, households, firms, countries…) at a point in time.
- Time-series data comprises observations on a set of variables for a particular unit (firm, country,…) across different points in time.
- Panel data combines the two, such that we have observations on a set of variables across a sample of units, and also across different points in time.



xwaveid	wave	cpi	ls	fs	age	healthcond
100001	1	1.42148	8	6	49	1
100001	2	1.38251	8	5	50	1
100001	3	1.34733	10	7	51	1
100001	4	1.3139	8	7	52	1
100002	1	1.42148	7	0	48	1
100002	2	1.38251	7	3	49	0
100002	3	1.34733	7	2	50	0
100002	4	1.3139	7	6	51	0
100003	1	1.42148	10	7	48	0
100003	2	1.38251	7	5	49	1
100003	3	1.34733	7	3	50	0
100003	4	1.3139	8	6	51	0



11

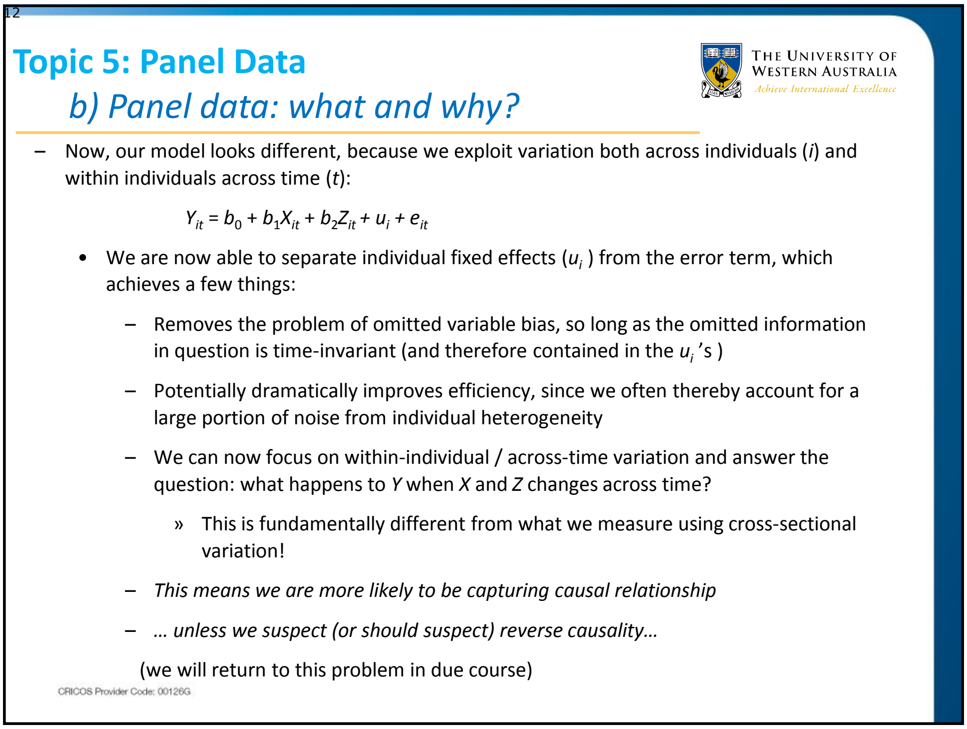
Topic 5: Panel Data

b) Panel data: what and why?

The University of Western Australia
Achieve International Excellence

- Panel data allow us to look for patterns both *within* and *across* individuals, firms, countries, etc, over a period of time (other units of analysis can be used in place of time, but we do not consider these contexts here)**.**
  - ➢ That is, we exploit both variation observed *across* individuals and variation observed *within* individuals across time.
  - ➢ Why can this be advantageous?

CRICOS Provider Code: 00126G



12

Topic 5: Panel Data

b) Panel data: what and why?

The University of Western Australia
Achieve International Excellence

– Now, our model looks different, because we exploit variation both across individuals (*i*) and within individuals across time (*t*):

$$Y_{it} = b_0 + b_1X_{it} + b_2Z_{it} + u_i + e_{it}$$

- We are now able to separate individual fixed effects (u_i) from the error term, which achieves a few things:
  - – Removes the problem of omitted variable bias, so long as the omitted information in question is time-invariant (and therefore contained in the u_i ’s)
  - – Potentially dramatically improves efficiency, since we often thereby account for a large portion of noise from individual heterogeneity
  - – We can now focus on within-individual / across-time variation and answer the question: what happens to *Y* when *X* and *Z* changes across time?
    - » This is fundamentally different from what we measure using cross-sectional variation!
  - – *This means we are more likely to be capturing causal relationship*
  - – *… unless we suspect (or should suspect) reverse causality…*

(we will return to this problem in due course)

CRICOS Provider Code: 00126G

13

Topic 5: Panel Data
c) Panel data models

- With panel data, we can analyse four different types of variation:

  1. Pooled panel models:
     - Simply pool the data = heap it all in a big pile and apply standard OLS as we would with cross-sectional data.
     - We are now effectively adding lots of observations, which increases our sample size, which in turn should provide more efficient estimates (though these estimates may be biased or inconsistent etc, if the classical assumptions underpinning OLS are violated)

  2. Between-group models:
     - Calculate mean values for each individual, such that each individual has got one value rather than several time-specific values.
     - Apply OLS to these means – effectively a cross-sectional model with individual means
     - Can be useful when measurement error is an issue, for example, but not often used.

14

Topic 5: Panel Data
c) Panel data models

3. Random-effects (RE) panel models:
   - Combines variation observed across and within individuals to improve on efficiency
   - Can be useful when we have very few time-observations (narrow panel) but desperately need to exploit all available data to the max.
   - But implies some troublesome assumptions
   - Tends to be used only when it’s the best available alternative
   - There are version of the RE model which can circumvent some of the usual problems..

4. Fixed-effects (FE) panel models:
   - Separates out individual fixed effects (u_i *‘s*) and considers only the variation observed within individuals, across time, to estimate model parameters.
   - Means we can relax the assumption that the u_i ‘s are uncorrelated with the explanatory variables, which circumvents many of the endogeneity issues we then otherwise would have to contend with.
   - So: This is the gold-standard for estimating causality in the absence of fancy IV techniques, *BUT*:
   - If we suspect reverse causality we still have issues…

Topic 5: Panel Data

c) Panel data models

From Kennedy p. 283:

A case where RE model would be fine, because it would be consistent with FE

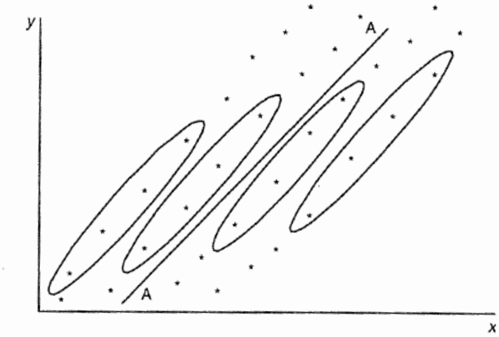


Figure 18.1 Panel data showing four observations on each of four individuals.

Topic 5: Panel Data

c) Panel data models

From Kennedy p. 285:

A case where RE model would yield biased estimates, and is not consistent with FE estimates:

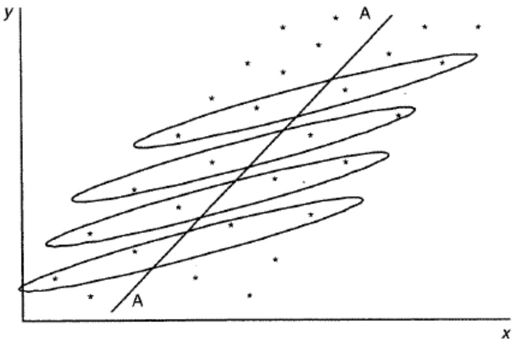
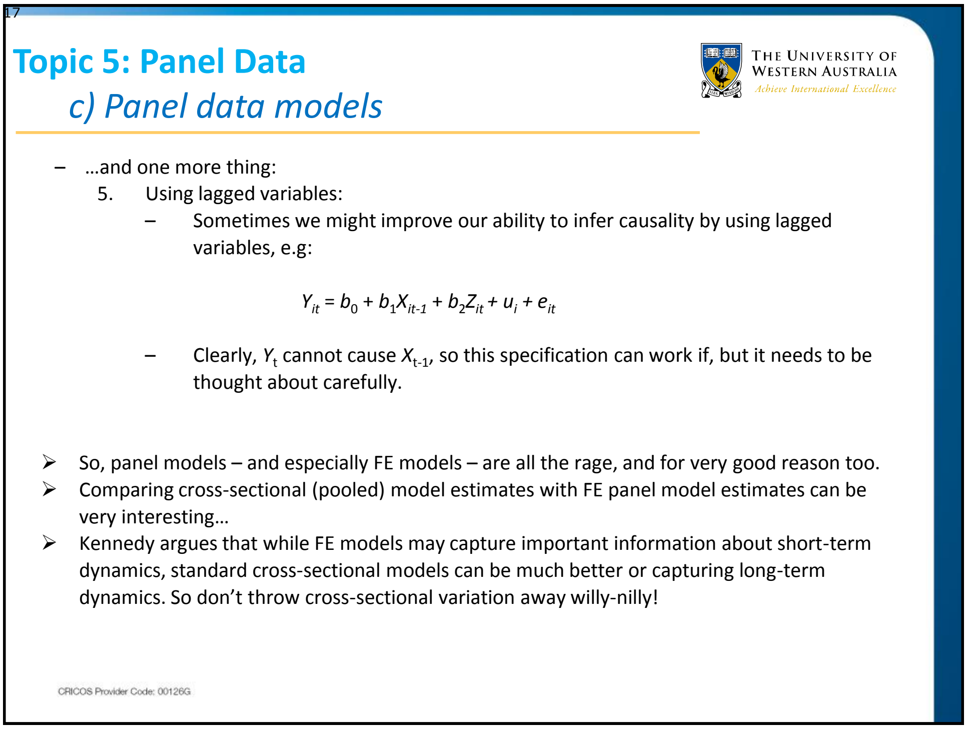


Figure 18.2 Panel data showing four observations on each of four individuals, with positive correlation between x and the intercept.



Topic 5: Panel Data

c) Panel data models

- …and one more thing:
  5. Using lagged variables:
     - Sometimes we might improve our ability to infer causality by using lagged variables, e.g:

$$Y_{it} = b_0 + b_1X_{it-1} + b_2Z_{it} + u_i + e_{it}$$

     - Clearly, Y_t cannot cause X_{t-1}, so this specification can work if, but it needs to be thought about carefully.

- ➢ So, panel models – and especially FE models – are all the rage, and for very good reason too.
- ➢ Comparing cross-sectional (pooled) model estimates with FE panel model estimates can be very interesting…
- ➢ Kennedy argues that while FE models may capture important information about short-term dynamics, standard cross-sectional models can be much better or capturing long-term dynamics. So don’t throw cross-sectional variation away willy-nilly!