THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

EFMD EQUIS ACCREDITED

ECON2271

Business Econometrics

Or: Practical Econometrics for Beginners

Week 12: Panel Data Applications;

Research Design

Topic 5: Panel Data; Topic 6: Research design

Agenda and learning outcomes

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Topic 5: Panel Data and Panel Models

a) Problems with cross-sectional data

b) Panel data: what and why?

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 6: Research design

- How to avoid FUQs
- The ultimate research design: RCT

❖ Text reference: Angrist, J.D. & Pischke, J-S. (2008) *Mostly Harmless Econometrics: An empiricist's companion.* Chapters 1&2

Topic 5: Panel Data
Recap

- A panel model looks different to purely cross-sectional and purely time-series models, 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!

Topic 5: Panel Data
Recap

The University of Western Australia
Achieve International Excellence

- With panel data, we can analyse four different types of variation:
  1. **Pooled panel models:** Treats all observations as single unrelated observations (even if many are from same individual)
  2. **Between-group (BE) models:** Calculates mean obs for each group (e.g. individual) and estimates cross-sectional model based on group means
  3. **Random-effects (RE) panel model:** Combines variation observed across and within individuals to improve on efficiency, but problematic assumption that within and across group associations are the same.
  4. **Fixed-effects (FE) panel model:** 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.

- Because the assumptions underpinning RE models are usually treated as questionable, most studies compare pooled and FE panel model estimates. Hence, we will focus on how to contrast and interpret these.

CRICOS Provider Code: 00126G

Topic 5: Panel Data
Applications

- **Example 1**: A panel model for physical health (*ph*)
  - We use a panel based on the first 14 waves of the HILDA survey.
  - First, we will use the *xtset and xtdescribe* commands to describe the data.
    - This is important, especially for FE panel models: if we are going to use variation observed within individuals across time, we'd better make sure we actually have a decent amount of such variation to work with.
    - Panels can be balanced or unbalanced. A balanced panel consists of units (here, individuals) for whom we have complete information. An unbalanced panel includes complete and incomplete "rows".
  - Second, we will estimate the model treating the data as pooled, and evaluate these estimates.
  - Third, we will estimate a FE panel model, and contrast and evaluate these estimates.

Topic 5: Panel Data
Applications

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

```
. xtset xwaveid wave
       panel variable:  xwaveid (unbalanced)
        time variable:  wave, 1 to 14, but with gaps
                delta:  1 unit

. xtdescribe

 xwaveid:  100001, 100002, ..., 1401168                      n =      20666
    wave:  1, 2, ..., 14                                     T =         14
           Delta(wave) = 1 unit
           Span(wave)  = 14 periods
           (xwaveid*wave uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         4        12      14      14

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+----------------
     4139     20.03   20.03 |  11111111111111
     2429     11.75   31.78 |  ..........1111
     1065      5.15   36.94 |  1.............
      721      3.49   40.42 |  .............1
      662      3.20   43.63 |  11............
      556      2.69   46.32 |  ............11
      492      2.38   48.70 |  ...........111
      450      2.18   50.88 |  111...........
      353      1.71   52.58 |  .........11111
     9799     47.42  100.00 | (other patterns)
 ---------------------------+----------------
    20666    100.00         |  XXXXXXXXXXXXXX
```

- ➢ This list produces the most commonly observed patterns
- ➢ 20% of individuals have complete information for waves 1-14
- ➢ Clearly, this is an unbalanced panel
- ➢ This is not considered a problem here, but it may be in other contexts.

Topic 5: Panel Data
Applications

```
. reg ph lnIncome female age agesq partnered separated divorced unemployed nonpartic
> ipant highschool trade undergraduate postgraduate

      Source |       SS           df       MS      Number of obs   =   117,629
-------------+----------------------------------   F(13, 117615)   =   1734.76
       Model |  8234811.73        13  633447.056   Prob > F        =    0.0000
    Residual |  42946968.2   117,615  365.148733   R-squared       =    0.1609
-------------+----------------------------------   Adj R-squared   =    0.1608
       Total |  51181779.9   117,628  435.115618   Root MSE        =    19.109

-------------------------------------------------------------------------------
           ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
     lnIncome |   1.782458    .063132    28.23   0.000      1.65872    1.906195
       female |   .6781491   .1156187     5.87   0.000     .4515382    .9047599
          age |  -.6906517   .0450884   -15.32   0.000    -.7790243   -.6022791
        agesq |   .0045757   .0004977     9.19   0.000     .0036002    .0055512
    partnered |   1.869205   .1756006    10.64   0.000     1.525031     2.21338
    separated |  -2.314194   .3159926    -7.32   0.000    -2.933534   -1.694854
     divorced |  -2.672584   .2717262    -9.84   0.000    -3.205163   -2.140005
   unemployed |  -6.157563   .3438218   -17.91   0.000    -6.831448   -5.483678
nonparticipant |   -13.9925   .1492823   -93.73   0.000    -14.28509   -13.69991
   highschool |   2.930104   .1971012    14.87   0.000     2.543789    3.316419
        trade |   1.290604   .1484063     8.70   0.000     .9997304    1.581478
undergraduate |   3.489805   .1651638    21.13   0.000     3.166086    3.813523
 postgraduate |   4.712929   .2743907    17.18   0.000     4.175127     5.25073
        _cons |   78.13877   1.170179    66.78   0.000     75.84524     80.4323
-------------------------------------------------------------------------------
```

Topic 5: Panel Data
Applications

- Everything is very significant. No wonder, with N being so huge!
- The largest and most significant coefficient is for *nonparticipant*:
  - People who do not participate in the labour market report physical health scores that are, on average, 14 points lower than people who employed, holding all other included variables constant.
  - But does this mean that exiting (entering) the workforce is associated with a deterioration (improvement) in health?
  - Even if we are controlling for a whole heap of potentially confounding variables, this difference could capture differences in some other variable we can't observe. Hence, we have to be careful implying anything other than cross-sectional association here.
- Look at the coefficients for marital status:
  - People who are partnered are healthier than others, and people who are separated and divorced are less healthy (the reference is clearly everybody else, i.e. not partnered, separated or divorced).
  - But does this mean that *becoming* partnered is associated with an *improvement* in health? Or that health deteriorates when going through separation and divorce?
- We could look at all the coefficients in a similar way, but let's just focus on these for now.

```
. xtreg ph lnIncome female age agesq partnered separated divorced unemployed nonpart
> icipant highschool trade undergraduate postgraduate, fe
note: female omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =    117,629
Group variable: xwaveid                         Number of groups   =     18,175

R-sq:                                           Obs per group:
     within  = 0.0271                                         min =          1
     between = 0.1062                                         avg =        6.5
     overall = 0.0853                                         max =         14

                                                F(12,99442)        =     230.39
corr(u_i, Xb)  = 0.0567                         Prob > F           =     0.0000

------------------------------------------------------------------------------
          ph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnIncome |   .2099206   .0505869     4.15   0.000     .1107708    .3090704
      female |          0  (omitted)
         age |   .0605936   .0555521     1.09   0.275    -.0482877    .1694749
       agesq |   -.005189   .0005962    -8.70   0.000    -.0063576   -.0040204
   partnered |  -.2598361   .2207967    -1.18   0.239    -.6925949    .1729226
   separated |   .6416996   .2855794     2.25   0.025     .0819674    1.201432
    divorced |  -.2305519   .3406671    -0.68   0.499    -.8982552    .4371514
  unemployed |   -.427514   .2644037    -1.62   0.106    -.9457421    .0907141
nonparticipant |  -4.154164   .1503503   -27.63   0.000    -4.448848   -3.859479
  highschool |   -.850809   .6142323    -1.39   0.166    -2.054697    .3530788
       trade |  -1.187832   .4041256    -2.94   0.003    -1.979914   -.3957511
undergraduate |  -2.749542   .6750693    -4.07   0.000    -4.072669   -1.426414
postgraduate |  -2.232844   .8491254    -2.63   0.009    -3.897119   -.5685685
       _cons |   85.42735   1.359018    62.86   0.000     82.76369    88.09101
```

10

Topic 5: Panel Data

Applications

- Not everything is significant!
- The coefficient for *female* says "omitted". Why??
- The coefficient for *lnIncome* is MUCH smaller. Why? What does this mean?
- Look at the coefficient for *nonparticipant*:
  - Exiting (entering) the workforce is associated with a deterioration (improvement) in health of 4.15 points on the health index! Much smaller than the coefficient from the pooled model, but still some significant within-individual association.
  - But does it mean that exiting the workforce is detrimental to health (or that entering will lead to improvements)? Can we infer a causal effect from this? Could there be something else going on?
- Look at the coefficients for marital status:
  - It turns out that becoming partnered is *not* associated with any significant change in health. So, the positive association in the pooled model is just from the fact that people who are partnered tend to be healthier than people who are not, for whatever reason.
  - Becoming separated is actually associated with an improvement in health! And it's statistically significant (just)! Could mean that people do get healthier when they go through separation. What do you think?
  - No statistically significant association between becoming divorced and health.

11

Topic 5: Panel Data
Applications

- **Example 2**: Does psychological distress cause risky health behaviours?
  - Research paper by Dan Hoang, Inga Kristoffersen and Ian Li.
  - First, we use a standard FE panel model to evaluate what happens to risky health behaviours (HB) when we observe changes in psychological distress (K10), holding a range of other potential confounding variables (X) constant.

$$HB_{it} = \alpha_0 + \alpha_1 K10_{it} + \alpha_2 X_{it} + u_i + \varepsilon_{it} \qquad (1)$$

However, we expect a fair amount of bias from reverse causality. In fact, most prior work looks at how health behaviours affect mental health, rather than the other way around, finding plenty of evidence of positive associations.

  - We want to look at how mental health affects health behaviours, which means we have to try and find a variable which is correlated with psychological distress but not with health behaviours. We find that a negative financial shock works well.
  - We then use this variable to create an artificial (instrumental) variable (an IV) which captures only the "purely exogenous" variation in K10. This IV is then used instead of K10 to estimate another version of equation (1).
  - This is essentially incorporating a 2SLS procedure into the FE panel model.

12

Topic 5: Panel Data
Applications

The University of Western Australia
Achieve International Excellence

- **Example 2**: Does psychological distress cause risky health behaviours?
  - The first-stage regression (using FE):

$$K10_{it} = \theta_0 + \theta_1 V_{it} + \theta_2 X_{it} + u_i + \varepsilon_{it} \qquad (3)$$

  - The second stage regression (using the IV = $\widehat{K10}$)

$$HB_{it} = \mu_0 + \mu_1 \widehat{K10}_{it} + \mu_2 X_{it} + u_i + \varepsilon_{it} \quad (4)$$

13

Topic 5: Panel Data
Applications

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **Example 2**: Does psychological distress cause risky health behaviours?

Dependent Variable:	Gender	Pooled (across) model K10 coeff	FE (within) model K10 coeff	FE model with IV K10 coeff	Pooled model with IV K10 coeff	FE IV with imputed data K10 coeff
Unhealthy eating index^1	All	0.0204***	0.0437***	-0.0099	0.0189***	0.0477
	Women	0.0244***	0.0472***	-0.0280	0.0092	-0.0332
	Men	0.0201***	0.0426***	0.0033	0.0281***	0.1139
Does not eat vegetables daily	All	0.0070***	0.0305***			
	Women	0.0080***	0.0269**			
	Men	0.0068***	0.0358**			
Does not eat fruit daily	All	0.0066***	0.0284***			
	Women	0.0087***	0.0377***			
	Men	0.0056***	0.0205			
Eats fried potatoes weekly or more	All	0.0067***	0.0317***			
	Women	0.0076***	0.0322**			
	Men	0.0075***	0.0318**			

*** denotes $p < 0.01$
** denotes $p < 0.05$
* denotes $p < 0.1$

14

Topic 5: Panel Data
Applications

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Dependent Variable:	Gender	Pooled (across) model K10 coeff	FE (within) model K10 coeff	FE model with IV K10 coeff	Pooled model with IV K10 coeff	FE IV with imputed data K10 coeff
Unhealthy lifestyle index^2	All	0.0436***	0.0755***	0.0313**	0.0650***	0.2722***
	Women	0.0473***	0.0893***	0.0269	0.0554***	0.2910***
	Men	0.0423***	0.0606***	0.0365*	0.0761***	0.3129***
Binge drinking	All	0.0070***	0.0232***			
	Women	0.0097***	0.0310***			
	Men	0.0054***	0.0115			
Current smoker	All	0.0103***	0.0207***			
	Women	0.0105***	0.0250***			
	Men	0.0107***	0.0155**			
Physical activity < 3 times a week	All	0.0115***	0.1018***			
	Women	0.0101***	0.1039***			
	Men	0.0125***	0.1040***			
Does not eat breakfast regularly	All	0.0123***	0.0342***			
	Women	0.0136***	0.0363***			
	Men	0.0118***	0.0303**			

*** denotes $p < 0.01$
** denotes $p < 0.05$
* denotes $p < 0.1$

15

Topic 5: Panel Data
Applications

- **Example 2**: Does psychological distress cause risky health behaviours?

…Soon to be published in a very good journal! (we hope…)

16

Topic 6: Research Design
Good research design

What is the question??

- What is the causal relationship of interest?
- What experiment could ideally be used to capture the causal effect of interest?
- What is the identification strategy applied?
- What is the mode of statistical inference?

The experimental ideal:

- Randomised Controlled Trials! (RCTs)
- Not always possible, but we try to emulate this design as closely as we can.
- The most credible and influential research designs use random assignment.

Topic 6: Research Design

Good research design

The selection problem: Do hospitals make people healthier? (A&P p. 10)

Since those admitted to the hospital get many valuable services, the answer to the hospital-effectiveness question still seems likely to be "yes". But will the data back this up? The natural approach for an empirically-minded person is to compare the health status of those who have been to the hospital to the health of those who have not. The National Health Interview Survey (NHIS) contains the information needed to make this comparison. Specifically, it includes a question "During the past 12 months, was the respondent a patient in a hospital overnight?" which we can use to identify recent hospital visitors. The NHIS also asks "Would you say your health in general is excellent, very good, good, fair, poor?" The following table displays the mean health status (assigning a 1 to excellent health and a 5 to poor health) among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

The difference in the means is 0.71, a large and highly significant contrast in favor of the *non-hospitalized*, with a t-statistic of 58.9.

Topic 6: Research Design

Good research design

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

The selection problem & revelations from random assignment

How relevant is our hospitalization allegory? Experiments often reveal things that are not what they seem on the basis of naive comparisons alone. A recent example from medicine is the evaluation of hormone replacement therapy (HRT). This is a medical intervention that was recommended for middle-aged women to reduce menopausal symptoms. Evidence from the Nurses Health Study, a large and influential non-experimental survey of nurses, showed better health among the HRT users. In contrast, the results of a recently completed randomized trial shows few benefits of HRT. What's worse, the randomized trial revealed serious side effects that were not apparent in the non-experimental data (see, e.g., Women's Health Initiative [WHI], Hsia, *et al.*, 2006).

Topic 6: Research Design

Good research design

The selection problem & revelations from random assignment

An iconic example from our own field of labor economics is the evaluation of government-subsidized training programs. These are programs that provide a combination of classroom instruction and on-the-job training for groups of disadvantaged workers such as the long-term unemployed, drug addicts, and ex-offenders. The idea is to increase employment and earnings. Paradoxically, studies based on non-experimental comparisons of participants and non-participants often show that after training, the trainees earn less than plausible comparison groups (see, e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Lalonde 1995). Here too, selection bias is a natural concern since subsidized training programs are meant to serve men and women with low earnings potential. Not surprisingly, therefore, simple comparisons of program participants with non-participants often show lower earnings for the participants. In contrast, evidence from randomized evaluations of training programs generate mostly positive effects (see, e.g., Lalonde, 1986; Orr, et al, 1996).

Topic 6: Research Design

Fundamentally Unidentified Questions

Fundamentally Unidentified Questions (FUQs)

Research questions that cannot be answered by any experiment are FUQ'd: Fundamentally Unidentified Questions. What exactly does a FUQ'd question look like? At first blush, questions about the causal effect of race or gender seems like good candidates because these things are hard to manipulate in isolation ("imagine your chromosomes were switched at birth"). On the other hand, the issue economists care most about in the realm of race and sex, labor market discrimination, turns on whether someone treats you differently because they *believe* you to be black or white, male or female. The notion of a counterfactual world where men are perceived as women or vice versa has a long history and does not require Douglas-Adams-style outlandishness to entertain (Rosalind disguised as Ganymede fools everyone in Shakespeare's *As You Like It*). The idea of changing race is similarly near-fetched: In *The Human Stain*, Philip Roth imagines the world of Coleman Silk, a black Literature professor who passes as white in professional life. Labor economists imagine this sort of thing all the time. Sometimes we even construct such scenarios for the advancement of science, as in audit studies involving fake job applicants and resumes.[2]

21

Topic 6: Research Design

Fundamentally Unidentified Questions

Fundamentally Unidentified Questions (FUQs)

A little imagination goes a long way when it comes to research design, but imagination cannot solve every problem. Suppose that we are interested in whether children do better in school by virtue of having started school a little older. Maybe the 7-year-old brain is better prepared for learning than the 6 year old brain. This question has a policy angle coming from the fact that, in an effort to boost test scores, some school districts are now entertaining older start-ages (to the chagrin of many working mothers). To assess the effects of delayed school entry on learning, we might randomly select some kids to start kindergarten at age 6, while others start at age 5, as is still typical. We are interested in whether those held back learn more in school, as evidenced by their elementary school test scores. To be concrete, say we look at test scores in first grade.

22

Topic 6: Research Design

Fundamentally Unidentified Questions

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Fundamentally Unidentified Questions (FUQs)

The problem with this question - the effects of start age on first grade test scores - is that the group that started school at age 7 is . . . older. And older kids tend to do better on tests, a pure maturation effect. Now, it might seem we can fix this by holding age constant instead of grade. Suppose we test those who started at age 6 in second grade and those who started at age 7 in first grade so everybody is tested at age 7. But the first group has spent more time in school; a fact that raises achievement if school is worth anything. There is no way to disentangle the start-age effect from maturation and time-in-school effects as long as kids are still in school. The problem here is that start age equals current age minus time in school. This deterministic link disappears in a sample of adults, so we might hope to investigate whether changes in entry-age policies affected adult outcomes like earnings or highest grade completed. But the effect of start age on elementary school test scores is most likely FUQ'd.

CRICOS Provider Code: 00126G

23

Topic 6: Research Design

Good research design

But how common are randomly assigned experiments?

– Not common enough…

– They are often expensive and/or impractical, and we might not get to answer our question before we've waited a good number of years.

So what can we do in the absence of such experiments?

– Attempt to exploit cheaper and more readily available sources of variation.

– Try to find natural or quasi-experiments that mimic a randomized trial by changing the variable of interest while other factors are kept balance.

- E.g. a policy change that is rolled out exogenously, and (preferably) randomly across groups/areas or time.
- E.g. identifying a good instrumental variable which captures the variation we are interested in but not the variation we want to exclude.

– Sometimes, we can account for selection bias by applying statistical tools specifically designed to deal with this problem (in the absence of random assignment)

- E.g. Heckman selection model

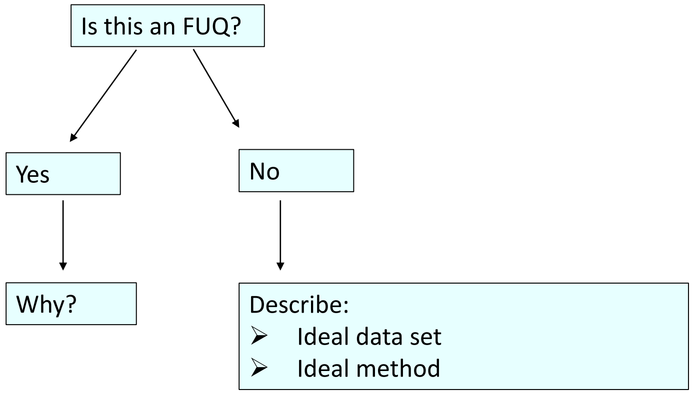
24

Topic 6: Research Design

Fundamentally Unidentified Questions

The University of Western Australia
Achieve International Excellence

Now you try: Come up with a question..



25

Topic 6: Research Design

Fundamentally Unidentified Questions

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Examples:

- If your parents split up when you were a child, are you more or less likely to form and keep stable relationships as an adult?
- Does pet-ownership improve mental health?
- Are left-handed children more creative than right-handed children?
- Are first-borns less happy than last-borns?
- Does better access to mental health professionals improve mental health outcomes?
- Does type of school (private or public) affect your labour market outcomes?
- Does type of school (co-ed or single-gender) affect risky sexual behavior?

26

Topic 6: Research Design

Good research design

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

- **In summary: Key tools for good econometric analysis include**
  1. Regression models designed to control for variables that may mask the causal effect of interest (we've covered that);
  2. Instrumental variables methods for the analysis of real and natural experiments (we've covered that a little – very briefly); and
  3. Differences-in-differences-type strategies that use repeated observations to control for unobserved omitted factors. (Not covered in this unit)