Panel Data

Introduction

- At the start of the course, I mentioned that there are three broad types of data:
  - Cross-section (n 'individuals' at one point in time)
  - Time series (T time periods for one 'individual')
  - **Panel data**
- Panel data (or longitudinal data) combines cross-sectional and time series data. (I realise it's a little unusual to study this before time series, but I have faith in you!)
- Panel data has n 'individuals' (people/companies/countries/etc.) sampled over several time periods (years/quarters/days/milliseconds/etc.).

Introduction

- We have typically been using $\{y_i, x_i\}$ where we have used an i subscript to denote the individual. (In time series we use a t subscript, as in $\{y_t, x_t\}$).
- You guessed it, we now have a double subscript using i and t, like $\{y_{it}, x_{it}\}$. So that $y_{10,3}$ denotes the 10th individual's observation in time period 3.
- We have skipped over time series for now because panel data follows much more closely the analysis of cross-sectional data. Also, panel data is great for identifying causal effects!

Panel Data

- It is possible (and in fact quite common) for panel data to have more than 2 dimensions. For example, we may have a sample of individuals (i) working at different companies (j) which reside in different countries (k) over different time periods (t).
- Panel data has become increasingly popular, in most part because of data availability and computing power.
- Some panel models can be very costly to estimate (in terms of computing power), and they are typically expensive to collect (in terms of time and money).

Famous Panel Datasets

- USA:
  - National Longitudinal Surveys of Labour Market Experience.
  - Panel Study of Income Dynamics
  - Health and Retirement Study
- Canada:
  - Survey of Labour Income Dynamics
- UK:
  - British Household Panel Survey
- A bunch of others from Germany, Belgium, France, Italy, China, Japan, Hungary, Netherlands, Russia, Switzerland, EU.

Types of Panel - Size

- **Short panels**: Large n, small T. (e.g. Household panels, Macroeconomic panels)
- **Long panels**: Small n, large T. (e.g. Grunfelds investment data: 10 large US manufacturing firms over time)
- **Huge panels**: Large n, large T. (e.g. Financial panels - stock prices for various companies every minute, or exchange rates)

Types of Panel - Structure

- **Balanced panels**: Every individual is observed in every time period of the data.
- **Unbalanced panels**: Some individuals are not observed in some time periods. This could be because, for example: some companies go bankrupt or are taken over; some people may die; some data is available further back for some countries; some people may not return the survey in some years. It is key to know if the data is 'missing at random' or not.
- **Pseudo panels**: A cross-section of different individuals are collected in each period (as the name suggests, not a real panel) but under certain assumptions can be used in a panel-like manner (cohort approach - group people based on observable characteristics).

Benefits of Panel Data

1. Large number of observations - self explanatory.
2. Controls for **unobserved heterogeneity** - we will discuss this in depth throughout the lecture. (Biggest benefit of panels)
3. Allows us to construct more complex models of behaviour - panels can relate an individual's experiences and behaviour at one point in time to other experiences and behaviours at another point in time. See the following example:

Example: Ben-Porath (1973)

- Suppose we have a cross section of women with a 50% average yearly labour force participation rate. This may be due to:

1. each woman having a 50% chance of being in the labour force in any given year, or
2. 50% of women work all the time and 50% don't at all.

- These two cases are very different and have different policy implications. A cross-section will not be able to tell us which is the true situation; a panel dataset can.

Another Example: The Glass Cliff

- Check out the 'freakonomics radio' podcast titled 'After the glass ceiling, a glass cliff' (Feb 2018). Just the first 7 minutes is enough.
- Notice the naive approach first. This doesn't identify a causal link, merely a correlation.
- More sophisticated techniques (including panel data) reveal the true story by uncovering the causal effect.

Panel Data - Setup

- Our basic model is

$$y_{it} = \alpha + x_{it}'\beta + u_{it}$$

  where x_{it} is a $(k \times 1)$ vector of regressors (not including a constant), $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

- For asymptotic analysis, it is generally assumed that $n \to \infty$ and T is fixed.

- Virtually all panel data models specify a structure on the error term u_{it}. Standard models are:

$$\begin{aligned} &\text{one-way effects :} && u_{it} = \mu_i + v_{it} \\ & && u_{it} = \lambda_t + v_{it} \\ &\text{two-way effects :} && u_{it} = \mu_i + \lambda_t + v_{it} \quad . \end{aligned}$$

- μ_i is the unobservable individual effect. λ_t is the unobservable time effect. v_{it} is the idiosyncratic component. We assume v_{it} are independent across individuals.

Panel Data - Independence

- Because we repeatedly observe the same units over time, it is usually no longer appropriate to assume that different observations are completely independent.
- For example, if we look at the demand for wine using a panel. The stuff in my unobservable term in period 1, $v_{\text{Luke},1}$, which contains, among other things, my preference for wine, is likely to contain a lot of the same things as $v_{\text{Luke},2}$, that is, my preferences in period 2.
- Now, these preferences may be 'time invariant', i.e. they will be contained in μ_i, and hence, not in v_{it}. However, my preferences can change over time, and so some degree of my 'taste' could be in v_{it}.
- This isn't a big issue, we'll see in the time-series part of the course what happens as a result, but it's useful to be aware of.

Unobserved Heterogeneity

- In general, the μ_i term is the key to narrowing in on causality.
- Suppose you have a panel of people, and you're interested in how education affects the likelihood of committing a crime. There are likely to be many things that you simply cannot capture: a person's general demeanor, how law-abiding their friends are, how susceptible they are to peer pressure, their individual experiences of crime when they were growing up...
- It is also likely that these things are correlated with the individual's level of education. So, by not controlling for them, we will have omitted variable bias and we cannot interpret our findings as causal.
- Econometricians have developed ways to estimate the causal effect β **without needing any information on** μ_i! And with no need for any instruments either!
- First, some assumptions:

Strict Exogeneity (One-way Model)

- A common assumption for the idiosyncratic component is

$$E\left[v_{it}\middle|\mu_i, x_{i1}, \ldots, x_{iT}\right] = 0 \text{ for } t = 1, \ldots, T.$$

- This says that the error term is mean-independent (hence also uncorrelated) with past, current, and future values of the regressors and the individual effect.
- Typically, this assumption is unlikely to hold.
- For example, consider a regression of a firm's output onto it's inputs (labour and capital). Do firms change their inputs in the next period $x_{i,(t+1)}$ based on a shock in this period v_{it}? Probably, yes.
- Strict exogeneity rules out any type of **feedback effect**.

Strict Exogeneity (One-way Model)

- Strict exogeneity also rules out any models with **lagged dependent variables**. This is where we regress y_{it} on $y_{i(t-1)}$. For example, you may want to forecast a particular company's stock price using the stock price from yesterday.
- We can often replace this strict exogeneity assumption with a weaker **sequential exogeneity** assumption:

$$E\left[v_{it}\middle|\mu_i, x_{i1}, \ldots, x_{it}\right] = 0 \text{ for } t = 1, \ldots, T.$$

- Notice that we only condition on current and past values of x, not future values.
- Also notice that nowhere have we restricted the relationship between μ and x. It is perfectly fine for them to be correlated; this, in fact, is the motivation for using panel data techniques - to account for the OVB.

Fixed Effects Estimator

- We are now going to show how to estimate $\hat{\beta}$ using the **fixed effects estimator**. This is also known as the **within estimator**.
- Recall, the one-way model (with just a single regressor for ease) is

$$y_{it} = \alpha + \mu_i + \beta x_{it} + v_{it}$$

- Technically, we have an issue with perfect multicolinearity. Since we have a μ for every observation, we can just drop the α, so each μ acts as the intercept for each individual,

$$y_{it} = \mu_i + \beta x_{it} + v_{it}.$$

- In most cases, we aren't interested in the μ_i, our parameter of interest is β. What we'll do is transform the model so that we **remove the** μ_i rather than try to estimate them.

Fixed Effects Estimator

- Consider the mean outcome for individual i:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} y_{it} &= \frac{1}{T}\sum_{t=1}^{T}\mu_i + \beta\frac{1}{T}\sum_{t=1}^{T} x_{it} + \frac{1}{T}\sum_{t=1}^{T} v_{it} \\ \bar{y}_i &= \mu_i + \beta\bar{x}_i + \bar{v}_i\end{aligned}$$

- Notice how the μ_i doesn't change. It is exactly the same for every time period, so its group average is just equal to itself.
- Of course, we can calculate this average for every i. Now, look what happens when we subtract the individual average from both sides of the regression:

$$\begin{aligned}y_{it} - \bar{y}_i &= (\mu_i + \beta x_{it} + v_{it}) - (\mu_i + \beta\bar{x}_i + \bar{v}_i) \\ &= \beta\left(x_{it} - \bar{x}_i\right) + \left(v_{it} - \bar{v}_i\right)\end{aligned}$$

- We've removed the μ_i! We still have the β! What's stopping us from regressing $(y_{it} - \bar{y}_i)$ on $(x_{it} - \bar{x}_i)$ using OLS? Nothing!

Example: Drink Driving (traffic1)

- We study the effect of two different drink driving laws on road deaths in the US. Open container laws (open) and administrative per se laws (admn).
- The first is self-explanatory. The second allows courts to suspend a driver's license if they are arrested for drink driving but before they are actually convicted.
- We have data on all 50 states (and Washington DC) for 2 years (not a very long panel!). We need to do a little bit of data work before we can run any regressions. Check out the R script on blackboard titled 'Panel.R'.
- Let's first run a simple **pooled OLS** model.

Example: Drink Driving

```
Call:
lm(formula = dthrte ~ admn + open, data = data_full)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13527 -0.42781 -0.06762  0.29307  1.96473

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.43527    0.09720  25.055   <2e-16 ***
admn         0.07216    0.12464   0.579    0.564
open        -0.10746    0.12708  -0.846    0.400
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.625 on 99 degrees of freedom
Multiple R-squared:  0.009499,  Adjusted R-squared:  -0.01051
F-statistic: 0.4747 on 2 and 99 DF,  p-value: 0.6235
```

Example: Drink Driving

- Not very convincing effects. We would hope that each of them has a significant negative coefficient.
- However, states decide whether they need such laws. Therefore, the presence of these laws is likely to be related to the average drink driving fatality rate in recent years.
- That is, those that have the laws are likely to be the ones that have high fatality rates!
- A fixed effects regression should be able to control for these state effects.

Example: Drink Driving

```
# Now a fixed effects model
library(plm)

PLM = plm(data=data_full, dthrte ~ admn + open,
          effect="individual",
          index="state")
summary(PLM)
```

Example: Drink Driving

```
Oneway (individual) effect Within Model

Call:
plm(formula = dthrte ~ admn + open, data = data_full, effect = "individual",
    index = "state")

Balanced Panel: n = 51, T = 2, N = 102

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.70000 -0.24871  0.00000  0.24871  0.70000

Coefficients:
     Estimate Std. Error t-value Pr(>|t|)
admn -0.51034    0.18527 -2.7546 0.008225 **
open -0.79655    0.33826 -2.3549 0.022580 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    10.79
Residual Sum of Squares: 8.1295
R-Squared:      0.24657
Adj. R-Squared: -0.55298
F-statistic: 8.01806 on 2 and 49 DF, p-value: 0.00097166
```

Example: Drink Driving

- This looks far more convincing. Both laws have a significant effect in reducing drink driving fatalities.
- Open container laws have a stronger effect than administrative per se laws but they're both pretty decent. The outcome variable is the number of drink related driving fatalities per 100 million miles driven per year. The number of miles driven in an average state is 40 billion (i.e. 40 thousand million miles). So the coefficient on *open* indicates that by introducing this law, an average state would save 0.8*400 = 320 people. (The average number of deaths in a state is 920)
- Ignore the R^2 and the adjusted R^2. They're not really meaningful in panel regressions (they shouldn't really be reported by R).

Finite Sample Properties of $\hat{\beta}$

- Unbiasedness: actually pretty straightforward, just redefine your outcome variable as $(y_{it} - \bar{y}_i)$ and your regressor as $(x_{it} - \bar{x}_i)$. You can then just recycle the same proof as used for OLS.
- Recall, for unbiasedness, we need to condition on all x_{it}. (We'll talk about consistency in a minute.)
- Variance: in the same way, we can recycle our proof from the OLS case. There, we showed that for the single regressor model (with no intercept)

$$Var\left(\hat{\beta}\middle|X\right) = \frac{\sigma^2}{\sum x_i^2}.$$

- Therefore, unsurprisingly, for the fixed effects model we have

$$Var\left(\hat{\beta}\middle|X\right) = \frac{\sigma^2}{\sum (x_{it} - \bar{x}_i)^2}.$$

Finding $\hat{\mu}_i$ (the Individual Effects)

- Technically, we use the Frisch-Waugh-Lovell Theorem to derive the $\hat{\mu}$. However, we'll go straight to the formula because it's fairly intuitive anyway.
- It can be shown that

$$\hat{\mu}_i = \bar{y}_i - \bar{x}_i' \hat{\beta}$$

- This is very similar to the intercept using the simple OLS estimator:

$$\hat{\alpha} = \bar{y} - \bar{x}' \hat{\beta}.$$

- So all we've actually done is calculate a separate intercept for each group.
- This highlights the fact that we haven't calculated just the k slope parameters in $\hat{\beta}$, but also n individual effects. So our degrees of freedom (for testing, etc.) are $nT - n - k$. So we ideally want T to be fairly large.

Asymptotic Properties of $\hat{\beta}$ and $\hat{\mu}$

- Recall, we assume $n \to \infty$ and T is fixed.
- $\hat{\beta}$ and $\hat{\mu}$ behave very different asymptotically.
- We can show $\hat{\beta}$ is asymptotically Normal and consistent. Recall that for OLS we only required $E[u_i x_i] = 0$. Now, we require $E[u_{it}(x_{it} - \bar{x}_i)] = 0$.
- Because of the $\bar{x}_i$ term, we require $E[u_{it} x_{is}] = 0 \ \forall s$, i.e we need strict exogeneity for consistency to hold.
- $\hat{\mu}$ is inconsistent (although it is unbiased under certain assumptions). The inconsistency is because the number of things to estimate goes to infinity at the same rate as the number of observations. If n was fixed and $T \to \infty$ then $\hat{\mu}$ would be consistent.

Two-Way Model

- We have seen how to include individual fixed effects for each i. We can do the same thing in a one way model for t, but what about including both **individual and time effects**?
- Our model now becomes

$$y_{it} = \alpha + \mu_i + \lambda_t + x'_{it}\beta + v_{it}$$

- Again, there's perfect multicollinearity here. We have an intercept, we also have that the sum of the dummies corresponding to the μ_i is equal to one, and so too is the sum of the dummies corresponding to the λ_t.
- It's conventional to keep the α and drop one time effect and one individual effect. This means that each fixed effect is in relation to the time/individual which is left out (the baseline).

Two-Way Model

- Similarly to the one-way model, to obtain consistency, we assume strict exogeneity

$$E\left[v_{it}\middle|x_{i1},\ldots,x_{iT},\mu_i,\lambda_t\right]=0.$$

- Our estimator takes a similar form to the one way model. We simply run an OLS regression on

$$y_{it}-\bar{y}_i-\bar{y}_t+\bar{y}=\left(x_{it}-\bar{x}_i-\bar{x}_t+\bar{x}\right)'\beta+\left(v_{it}-\bar{v}_i-\bar{v}_t+\bar{v}\right)$$

- The reason we add back the $\bar{y}$, is because of the α term. To see how we get here, start with the model from the previous slide, then **subtract** $\bar{y}$ first and continue as we did before by subtracting $\bar{y}_i$ and $\bar{y}_t$. You'll end up with the above model where α and all of the fixed effects are removed.

Estimates of the Fixed Effects

- The estimates for μ_i, λ_t, and α are all very intuitive.

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \bar{x}'\hat{\beta} \\
\hat{\mu}_i &= (\bar{y}_i - \bar{y}) - (\bar{x}_i - \bar{x})'\,\hat{\beta} \\
\hat{\lambda}_t &= (\bar{y}_t - \bar{y}) - (\bar{x}_t - \bar{x})'\,\hat{\beta}
\end{aligned}$$

- $\hat{\alpha}$ is the same as our conventional OLS intercept. $\hat{\mu}_i$ and $\hat{\lambda}_t$ take very similar forms to the one-way model effects.
- While $\hat{\beta}$, $\hat{\alpha}$, and $\hat{\lambda}_t$ are all consistent, $\hat{\mu}_i$ remains inconsistent for the same reason as we discussed before.
- This problem of inconsistency due to an increasing number of parameters to estimate is known as **the incidental parameters problem**.

Testing for Poolability

- Just because we have panel data, does not mean that we have to use these fixed effects model. We could use a **pooled model**, this is where you just run a usual OLS regression and ignore the panel structure.
- To test if it is ok to pool the data, we simply test

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_{n-1} = \lambda_1 = \lambda_2 = \cdots = \lambda_{T-1}$$

- There are $n + T - 2$ restrictions here (recall that we drop one individual effect and one time effect to avoid perfect multicolinearity).
- Of course, we could test just for time effects, or just for individual effects if we wanted.
- Since the number of restrictions increases with the sample size, we cannot use any asymptotic test (Chi-Squared tests). Instead we must assume normality of the errors and use an F-Test.

Testing for Poolability

- To do this, we estimate the full fixed effects model, as well as the pooled model. Save the RSS for each model (this gives an indication of how well the model fits the data).
- Calculate the F-test as

$$F = \frac{(RSS_R - RSS_U)/(n+T-2)}{RSS_U/(nT-n-T-k+1)} \sim F_{n+T-2,nT-n-T-k+1}$$

- RSS_R is the residual sum of squares from the pooled (restricted) model. RSS_U is the restricted sum of squares from the fixed effects (unrestricted) model.
- The degrees of freedom on the top is the number of restrictions. The degrees of freedom on the bottom is the number of observations minus the number of parameters estimated in the unrestricted model (why?).

Example: Crime Rate in NC (crime4)

- We want to see whether probability of arrest, conviction, and a prison sentence effect the crime rate in North Carolina. We also see whether the average prison sentence and police per capita effect the probability of arrest.
- We have data on 90 counties in North Carolina over a seven year period in the 1980s.

```
PLM = plm(data=data, lcrmrte ~ lprbarr + lprbconv + lprbpris + avgsen + polpc,
          effect = "twoways",
          index = c("county", "year"))
summary(PLM)
```

Example: Crime Rate in NC

```
Coefficients:
           Estimate Std. Error  t-value  Pr(>|t|)
lprbarr  -0.3457585  0.0325689 -10.6162 < 2.2e-16 ***
lprbconv -0.2696145  0.0210493 -12.8087 < 2.2e-16 ***
lprbpris -0.1620330  0.0326637  -4.9606 9.484e-07 ***
avgsen    0.0027377  0.0027563   0.9933     0.321
polpc    59.6692877  3.8349672  15.5593 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    16.123
Residual Sum of Squares: 10.406
R-Squared:      0.3546
Adj. R-Squared: 0.23259
F-statistic: 58.1283 on 5 and 529 DF, p-value: < 2.22e-16
```

Example: Crime Rate in NC

- The signs of the coefficients are as we would expect... except for *polpc*.
- There are at least two possibilities.
  - The crime rate variable is calculated from reported crimes. It might be that, when there are additional police, more crimes are reported.
  - The police variable might be endogenous in the equation for other reasons: counties may enlarge the police force when they expect crime rates to increase. In either case, we cannot interpret our effect for police per capita as causal.
- So, it is likely that using panel data has not solved all of the endogeneity issues in this problem. It is possible to combine IV with panel data but we will not cover this in this course.

Example: Crime Rate in NC (Testing Poolability)

```
# Test for poolability

# Run restricted model
LM = lm(data=data, lcrmrte ~ lprbarr + lprbconv + lprbpris + avgsen + polpc)
summary(LM)

# Save residual sum of squares and degrees of freedom
RSS_R = sum(LM$residuals^2)
RSS_U = sum(PLM$residuals^2)
DofF1 = 90 + 7 - 2
DofF2 = 90*7 - 90 - 7 - 5 + 1

# Calculate F-test
F_test = ( (RSS_R - RSS_U) / DofF1 ) / ( RSS_U / DofF2 )

# Calculate p-value
pf(F_test, DofF1, DofF2, lower.tail = F)
```

Limitations of Fixed Effects Models

- Of course, we lose a lot of degrees of freedom when we estimate individual effects (and, to a lesser extent, time effects).
- But a bigger issue is that we are unable to estimate the effect of time-invariant variables.
- What does this mean? Take something like race. Your race doesn't change over time. When we subtract the individual means in order to remove the fixed effects, we also remove any time-invariant variables; they get lumped into the fixed effect.
- This can be very annoying when we want to look at race or gender bias!
- Also, as we have seen in the last example, fixed effects models may not remove all of the endogeneity in a model.

Summary

- We have seen what the structure of panel data looks like.
- We discussed the benefits of using panel data, in particular, the ability to control for unobserved heterogeneity.
- We briefly looked at the assumptions imposed in panel data contexts.
- We showed how to use the fixed effects estimator and some of its properties in one-way and two-way models.
- We looked at how to whether we need a fixed effects model by comparing to the pooled model.
- Finally, we have mentioned the limitations of fixed effects models