Acknowledgement of Country

I acknowledge the Traditional Owners and their custodianship of the lands on which we meet today.

On behalf of us all, I pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

Lecture 6: Advanced Panel Data methods

Julie Moschion
University of Queensland

Reading for lecture 6

- In the textbook (Wooldridge 2020): Chapter 14
- Mostly Harmless Econometrics (Angrist and Pischke): Chapter 5
- Additional Reading: Microeconometrics: Methods and Applications (Cameron and Trivedi): Chapter 21

Outline

- Panel data
- First Differences (FD) estimator
- Fixed Effects (FE) estimator
- Random Effects (RE) estimator
- Hausman Test

Panel data

- A panel dataset differs from a pooled cross section:
  - In a panel dataset, information from the **same individual units** are recorded at each point in time.
- Panel data can be assembled for a variety of economic units: countries, states, cities, postcodes, schools, firms, individuals...
- In microeconomic applications, we most often use data on many firms/individuals/... $(> 1,000)$ and relatively few time periods (< 50).
- With panel data, we cannot assume that observations are independent over time (we observe the same unit several times!). For each unit, errors are correlated over time leading to **serial correlation of errors.**
  - With individual-level data, the unobservable factors that affect an individual's wage will be present at each point in time, leading to correlations over time that we call **unobserved heterogeneity**.

Panel data

- Panel data analysis present many advantages.
- With multiple years of data we can solve omitted variable biases stemming from unobserved characteristics that do not change over time.
  - E.g. Control for unobserved family background variables (e.g. mother's ability) by comparing siblings within families.
- Multiple years of data enable the analysis of dynamic phenomena.
  - E.g. How does Australia's economic performance last year affect its performance this year?
- Relationships can be estimated more efficiently.
  - I.e. The increase in sample size increases the precision of estimators.

Panel data

- The best way **to store panel data** is to stack the units *i* on top of each other. In particular, the time periods for each unit should be grouped together, and stored in chronological order (from earliest period to the most recent). This is called the **"long"** storage format. It is the most common.
- When the data is "long", it is useful to have the data appropriately sorted by using `sort distid year` in Stata. To have a look at it, you can use `browse`.
- To set up your data as panel data in Stata, use `xtset distid year`. This tells Stata that the two dimensions of your panel data are the unit id and the year.
- Avoid sorting by year and then unit id. That would make the data set look more like independently pooled cross sections, and mask the panel structure.

Panel data

- Sometimes panel data sets (especially with two years) are organised **"wide"**.
- In this case, the data are stored as having only n records (rather than $2n$ like in the "long" format), with the variables from the different years organised in columns and given different suffixes (e.g. emp20 and emp21 for employment in 2020 and in 2021).
- Generally, this makes the data harder to work with, especially if there are more than two years.
- Stata has a command, `reshape`, that allows one to go from `wide` to `long`, and vice versa.

Panel data

- Long panel

	County	LandArea	NatAmenity	Year	College	Jobs
1	Autauga	599	4	1970	.064	6853
2	Autauga	599	4	1980	.121	11278
3	Autauga	599	4	1990	.145	11471
4	Autauga	599	4	2000	.180	16289
5	Baldwin	1578	4	1970	.065	19749
6	Baldwin	1578	4	1980	.121	27861
7	Baldwin	1578	4	1990	.168	40809
8	Baldwin	1578	4	2000	.231	70247
9	Barbour	891	4	1970	.073	9448
10	Barbour	891	4	1980	.092	9755
11	Barbour	891	4	1990	.118	12163
12	Barbour	891	4	2000	.109	15197
13	Bibb	625	3	1970	.042	3965
14	Bibb	625	3	1980	.049	4276
15	Bibb	625	3	1990	.047	5564
16	Bibb	625	3	2000	.071	6098
17	Blount	639	4	1970	.027	7587
18	Blount	639	4	1980	.053	9490
19	Blount	639	4	1990	.070	11811
20	Blount	639	4	2000	.096	16503

- Wide panel

	County	LandArea	NatAmenity	College1970	College1980	College1990	College2000	Jobs1970	Jobs1980	Jobs1990	Jobs2
1	Autauga	599	4	.064	.121	.145	.180	6853	11278	11471	16
2	Baldwin	1578	4	.065	.121	.168	.231	19749	27861	40809	70
3	Barbour	891	4	.073	.092	.118	.109	9448	9755	12163	15
4	Bibb	625	3	.042	.049	.047	.071	3965	4276	5564	6
5	Blount	639	4	.027	.053	.070	.096	7587	9490	11811	16

Example with 2 periods

- We have data on unemployment and crime rate for 46 cities for 1982 and 1987. Assume that t=1 (year=1982) and t=2 (year=1987). Using the 1987 cross-section we obtain:

$$\widehat{crmrte} = \underset{(20.76)}{128.38} - \underset{(3.42)}{4.16}\, unem$$
$$n = 46,\ R^2 = .033.$$

- If we interpret these estimates causally, it implies that an increase in the unemployment rate has no impact on the crime rate, but is that so?
- Really the only thing we learn from this cross-section is that in 1987, cities with more unemployment did not have more crime...
- But does that mean that if unemployment increases somewhere, the crime rate won't change? (causal)

Example with 2 periods

- Not quite...as we may have an omitted variable problem.
- For example, local unobservables such as age distribution, gender distribution, education levels, law enforcement efforts etc may contribute to explaining different unemployment and crime rates across cities.
- Solution 1: *Control for more factors* but many factors might be hard to control for.
- Solution 2: *Control for crmrte from a previous year* (in this case 1986) to adjust for the fact that different cities have different crime rates at baseline (autoregressive models, will not discuss them now).
- Solution 3: Use **panel data methods** to wash out the effect of the unobserved factors that are constant over time. In our previous example, these are all the things that are intrinsic to each different city, but varies across cities.

Panel data with 2 periods

- Let's look at this 2-period panel data model more formally.
- Each unit i is observed in two time periods: $t = 1, 2$. The units can be aggregated (schools or cities) or disaggregated (students or teachers).
- Consider the following model with a single explanatory variable:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 d_t + c_i + u_{it}, t = 1, 2.$$

- We observe x_{it} and y_{it} in both periods.
- The variable d_t is a constructed time dummy for the second time period: $d_t = 1$ if $t = 2$ and $d_t = 0$ if $t = 1$.
- In t=1 the **intercept** is β_0 and in t=2 the **intercept** is $\beta_0 + \beta_2$: allowing the intercept to change over time is important in most applications (in the previous example, the crime rates can change considerably over a five-year period.)

Panel data with 2 periods

- Our model with a single explanatory variable:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 d_t + c_i + u_{it}, t = 1, 2.$$

- **Time-varying component of error:** u_{it} is the unobserved idiosyncratic error or the time-varying error.
- **Time-invariant component of error:** c_i captures all unobserved, time-constant factors that affect y_{it} and in particular the heterogeneity across the units i. It is called the unobserved effect, unobserved heterogeneity or fixed effect (because it is fixed over time).
- We are interested in estimating β_1, the partial effect of x on y. Note that the model assumes this effect is constant over time (as in previous lectures).

Panel data with 2 periods

- Our model with a single explanatory variable:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 d_t + c_i + u_{it}, t = 1, 2.$$

- Suppose that we have data for 200 schools and 2 time periods.
- **Q1:** How many time dummies do you have when there are 2 time periods?
  - **Answer1**: 1.
- **Q2:** How would you interpret β_2?
  - **Answer2**: It is the difference in the mean of y in period 2 compared to period 1.
- **Q3:** How many variables are in c_i?
  - **Answer3**: 199.
- **Q4:** How would you interpret c_i?
  - **Answer4**: It is the difference in the mean of y for each school and the reference school.

Panel data with 2 periods

- How should we estimate the parameter of interest β_1 given that we have two years of panel data?
- One possibility is to just pool the two years and use OLS. To do this, we consider a composite error term $v_{it} = c_i + u_{it}$, $t = 1, 2$, and estimate:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 d_t + v_{it}, t = 1, 2.$$

- If we apply OLS, we will simply regress y on d and x (there are no unit fixed effects).
- Would this estimate have the desired properties?

Panel data with 2 periods

- **Two problems!**
- **Problem 1:** even if we assume random sampling across i, we cannot reasonably assume that the observations for i across $t = 1, 2$ are independent.
- The errors will be correlated because of c_i: $Cov(v_{i1}, v_{i2}) \neq 0$
- We call this **serial correlation** or **cluster correlation** (each unit i is a cluster with two time periods).
- The usual OLS standard errors will be wrong in this context.
- Just using heteroskedasticity-robust standard errors does not solve the problem.
- **Solution**: generate "cluster-robust" standard errors and test statistics is very easy with modern softwares.
- In Stata: `vce(cluster (clustervar))`.

Panel data with 2 periods

- **Problem 2:** the **exogeneity** assumption.
- In order for the pooled OLS to produce an unbiased and consistent estimator of β_1, we would have to assume that the composite error, v_{it}, is uncorrelated with x_{it}.
- Because $v_{it} = c_i + u_{it}$, we need both:

$$Cov(x_{it}, c_i) = 0; Cov(x_{it}, u_{it}) = 0$$

- Suppose we are willing to assume that the second holds.
- What about the first one?
- If $Cov(x_{it}, c_i) \neq 0$, it means that there are intrinsic features of units i that do not vary over time, which contribute to determining y_{it} and are correlated with x_{it}.
- This generates **heterogeneity bias** in pooled OLS.
- Solution?

Panel data with 2 periods

- Heterogeneity bias arising from $Cov(x_{it}, c_i) \neq 0$ can be managed in a number of ways:
  - First Difference estimation: pooled OLS but on the differences.
  - Fixed Effects estimation: pooled OLS but on the time-demeaned variables.
  - Random Effects estimation: pooled OLS but on the partially time-demeaned variables. It is a GLS procedure.
- FE and FD eliminate c_i, which can be quite powerful.

First-Difference estimator

- If the explanatory variable of interest x_{it} changes over time (at least for some units in the population), we can eliminate heterogeneity bias by differencing c_i away in the pooled OLS.
- Write the model for each time period starting with the last t (remember: d_t=0 in period 1, d_t=1 in period 2).

$$y_{i2} = \beta_0 + \beta_2 + \beta_1 x_{i2} + c_i + u_{i2}$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + c_i + u_{i1}$$

- Subtract time period 1 from time period 2 to get:

$$y_{i2} - y_{i1} = \beta_2 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

- If we define $\Delta y_i = y_{i2} - y_{i1}$, where Δ = *change over time*:

$$\Delta y_i = \beta_2 + \beta_1 \Delta x_i + \Delta u_i$$

- We can apply the OLS in the first-differenced model (c_i has been eliminated).
- Important: β_1 is the **original coefficient** we are interested in.

First-Difference estimator

- Differencing away the unobserved effect, c_i, is simple and can be very powerful for isolating causal effects.
- The OLS estimator of $\hat{\beta}_1$ is called the **First-Difference (FD) estimator** (with more than two time periods, other orders of differencing are possible; hence the qualifier "first".)
- However, for the **FD estimator to be consistent**, the following assumption has to hold:

$$Cov(x_{is}, u_{it}) = 0, \ \ for\ all\ s, t = 1, ..., T.$$

- We call this the **strong exogeneity assumption**.
- c_i is removed, so it cannot be a source of serial correlation. But the u_{it} may still be serially correlated (and heteroskedastic). So in Stata: Use the **"cluster-robust"** standard errors command.
- The same differencing strategy works if x_{it} is a binary program indicator. The differenced equation is the same.

Fixed Effects estimator

- We can also use the "fixed effects" or "within" transformation to remove c_i using the within i time averages.
- Essentially the "fixed effects" is a pooled OLS on the time-demeaned variables.
- In the simple model with only one x_{it}:

$$y_{it} = \beta_0 + \beta_1 x_{it} + c_i + u_{it} \tag{1}$$

- Average this equation across t to get:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + c_i + \bar{u}_i \tag{2}$$

- where $\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ is a "time average" for unit i. Similarly, for $\bar{x}_i$ and $\bar{u}_i$.
- Note that c_i doesn't vary over t, so: $\bar{c}_i = T^{-1} \sum_{t=1}^{T} c_i$ $=\mathrm{T}^{-1}\ Tc_i = c_i$.

Fixed Effects estimator

- Subtract (2) from (1) to obtain:

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i) \tag{3}$$

- We estimate (3) by OLS: this "time-demeaned" (or within) equation has become the estimating equation.
- However, as with FD, we interpret $\hat{\beta}_1$ as the estimator from the levels equation (1) and call it the **"Fixed Effects" (FE) estimator** or the **"within" estimator**.
- The important thing is that the unobserved effect, c_i, has disappeared.
- The intuition is that by removing the time averages that characterise the x and u of a unit, we have removed the time-invariant component of y within units.
- This is a very common technique in applied work!

Fixed Effects estimator

- As with FD, for the **FE estimator to be consistent**, the **strong exogeneity assumption** has to hold:

$$Cov(x_{is}, u_{it}) = 0, \ \ for\ all\ s, t = 1, ..., T.$$

  - The idiosyncratic error should be uncorrelated with each explanatory variable across all time periods.
- c_i is removed, so it cannot be a source of serial correlation. But the u_{it} may still be serially correlated (and heteroskedastic). So in Stata: use the **"cluster-robust"** standard errors command.
- Do not worry too much about R^2 with FE. The "within" R^2 is probably most informative (time-demeaned equation).

FD and FE in a nutshell

- **FE has an advantage over FD** when strict exogeneity fails: under reasonable assumptions with large T, **FE tends to be less biased than FD**.
  - If FD and FE are very different, it is a sign that strict exogeneity fails.
- However, a potential **drawback with FD and FE methods** is that any explanatory variable that is constant over time (x_i such as gender, or whether a city is located near a river...) gets swept away by the fixed effects or first differences transformation.
  - So the coefficient of those time-invariant variables cannot be estimated with these transformations: you won't get the effect of gender on the outcome.
  - But sometimes this doesn't matter so much and eliminating OVB is more important.

Random Effects Estimation

- Using the same model as before:

$$y_{it} = \mathbf{x}_{it}\beta + d_t + c_i + u_{it}$$

  where d_t represents a time dummy and c_i unit dummies.

- In a Random Effects (RE) model, we assume that the unobserved effect c_i is uncorrelated with each explanatory variable x_{it} in all time periods:
  $Cov(x_{it}, c_i) = 0, \ \textit{for all } t = 1, ..., T.$
- Unlike FD and FE, RE estimation leaves c_i in the error term.
- For consistency, RE also assume that the composite error term, $v_{it} = c_i + u_{it}$, is uncorrelated with the explanatory variables in all time periods:
  $Cov(x_{it}, v_{it}) = 0, \ \textit{for all } t = 1, ..., T.$
- In other words RE doesn't allow to correct for potential endogeneity!

Random Effects Estimation

- The "standard" RE assumptions also include that:

$$Cov(c_i, u_{it}) = 0 \text{ (not especially controversial)}$$

$$Var(u_{it}) = \sigma_u^2 \text{ for all } t \text{ (constant variance over time)}$$

$$Cov(u_{is}, u_{it}) = 0, t \neq s \text{ (no serial correlation)}$$

- The second and third assumptions often fail empirically but can be dealt with by computing clustered robust standard errors (as with FD and FE).
- RE accounts for the **serial correlation** over time in $v_{it} = c_i + u_{it}$ via a generalised least squares (GLS) procedure.

Random Effects Estimation

- For policy analysis, **RE is typically less convincing than FD or FE** as it doesn't deal with our main threat to identification: endogeneity.
- In some cases, it may be important to estimate coefficients for time-invariant explanatory variables (such as gender), which can drive the choice of an RE model. And there could be situations in which these time-invariant controls are sufficient to remove OVB making RE more convincing.
- **When RE is consistent, it is typically more efficient than FD or FE** – sometimes much more efficient (more precise estimates). This is because it uses all the variation in the data not just the within-unit variation.

Choosing between POLS, FD, FE, and RE

- We have covered four panel data estimators:
  - POLS, which is on the levels.
  - FD, which is POLS but on the differences over time.
  - FE, which is POLS on the time-demeaned variables.
  - RE, which is POLS on the partially time-demeaned variables.
- POLS on the levels is usually deficient, unless we include things like the lagged dependent variable y (not allowed in the other methods) and very convincing controls. But the possibility of unobserved variables is still looming, so many economists prefer models with unobserved effects.
- **FD versus FE**: if FD and FE are different in important ways, the strict exogeneity assumption may be violated and FE is preferred.

Choosing between POLS, FD, FE, and RE

- **RE versus FE**
  - Time-invariant variables drop out of FE estimation, so RE may be preferred in cases where this matters.
  - On the time-varying covariates:
    - When FE and RE are similar, it does not really matter which we choose.
    - When they differ a lot and in a statistically significant way, RE is likely to be inappropriate. Harder is when they differ by a lot practically but are not statistically different.
  - There is a way **to compare** the RE and FE estimators on explanatory variables that change across i and t: the **Hausman Test**.

Hausman Test

- We can use a formal statistical test to choose between RE and FE:
  - If $E(x_{it}c_i) = 0$ (H0), then $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ should be similar because both are consistent.
  - If $E(x_{it}c_i) \neq 0$ (H1), then $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ should be different because $\hat{\beta}_{FE}$ is consistent and $\hat{\beta}_{RE}$ is inconsistent.
- Therefore we can test whether $E(x_{it}c_i) = 0$ and hence choose between FE and RE estimators by testing whether $\hat{\beta}_{FE} = \hat{\beta}_{RE}$.
- The single parameter version of the test consists in examining the t-statistic:

$$t = \frac{\hat{\beta}_{FE} - \hat{\beta}_{RE}}{\sqrt{var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})}}$$

- When we reject the null hypothesis, then $E(x_{it}c_i) \neq 0$, and we prefer the FE estimator.

Hausman Test

- The Hausman test is performed conditional on proper specification of the underlying model.
- If we have omitted an important time-varying explanatory variable from both specifications, then we are comparing two inconsistent estimators of the population model.
- In Stata:

```
xtreg depvar indepvars1, fe
estimates store fe
xtreg depvar indepvars2, re
estimates store re
hausman fe re, sigmamore
```

Supplementary slides

FD and DiD

- In applications where x_{it} is a **dummy variable** with no treatment in the first period, the FD estimator is the same as applying OLS to:

$$\Delta y_i = \beta_2 + \beta_1 x_{i2} + \Delta u_i$$

where x_{i2} is the program participation in period 2 (zero or one). Note that $x_{i2} - x_{i1} = x_{i2}$ as $x_{i1} = 0$ for all i.

- The estimate of β_1 is the difference in means over time between the "treated" group and the "control" group (see lecture 5):

$$\hat{\beta}_{1,FD} = \overline{\Delta y}_{i,treat} - \overline{\Delta y}_{i,control}$$

- This is another **"difference-in-differences"** estimator.
- Here, unlike in the case of pooled cross sections, the differences are within the same unit: (i) differences in y between t2 and t1 for each i in the "treated" group are averaged; (ii) differences in y between t2 and t1 for each i in the "control" group are averaged. Then (i)-(ii) gives the DiD estimator.