Panel data and fixed effect estimator

A. Colin Cameron
Univ. of California, Davis

These slides are part of the set of slides
A. Colin Cameron, Introduction to Causal Methods
https://cameron.econ.ucdavis.edu/causal/

March 2023

Introduction

- These slides give an introductory example of the fixed effects estimator for panel data
  - the fixed effects estimator is a method for causal inference
- Here it is applied to panel data
  -
    - it also applies more generally to grouped data with group-specific effects
- Fixed effects causal inference relies crucially on the nontestable assumption that endogenous regressors are correlated only with a time-invariant (or group-invariant) component of the data.

- Separately the Stata file `panel.do` implements these methods
  - using dataset AED_NBA.DTA
- Data are from chapter 13.7 of A. Colin Cameron (2022) *Analysis of Economics Data: An Introduction to Econometrics* https://cameron.econ.ucdavis.edu/.
- Data originally from Edwin Bang (2012), "Collective Bargaining Agreements, Star Players, and Inequality in the NBA", Undergraduate Honors thesis, Dept. of Economics, U.C. Davis.

Outline

Panel data methods

- Panel or longitudinal data are data where the same individuals are observed at several points in time.
- The standard microeconometrics models and estimators have
  - data on many individuals
  - for at least two time periods (a short panel)
  - time observations that are equally spaced e.g. year or month.
- For individual i in time period t the pooled model sets
  - $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it}$, $i = 1, ..., n$, $t = 1, .., T$
  - the regressors can include controls for time trends.
- Statistical inference needs to use cluster-robust standard errors to control for likely time series correlation in the error term
  - cluster on individual assuming independence across individuals
  - cluster on group assuming independence across groups of individuals.

Fixed effects estimator

- The model is $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it}$.
- OLS is inconsistent if regressors are correlated with the error u_{it}.
- Fixed effects assume the regressors are correlated only with a part of the error that is constant over time
  - the crucial nontestable assumption.
- Assume that $u_{it} = \alpha_i + \varepsilon_{it}$ where
  - α_i is an individual-specific effect that may be correlated with regressors
    - e.g. in an earnings on schooling regression α_i may be unobserved ability
  - ε_{it} is an idiosyncratic part of the error that is uncorrelated with regressors.
- Define the individual-specific averages $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{x}_{2i}$, ..., $\bar{x}_{ki}$ and $\bar{\varepsilon}_i$ to be time averages for each individual.
- The fixed effects (or within) estimator is the OLS estimation in

$$(y_{it} - \bar{y}_i) = \beta_2(x_{2it} - \bar{x}_{2i}) + \cdots + \beta_k(x_{kit} - \bar{x}_{ki}) + \text{ error.}$$

Fixed effects estimator (continued)

- Proof: manipulation gives the within model or within model

$$\begin{aligned} y_{it} &= \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it} \\ \bar{y}_i &= \beta_1 + \beta_2 \bar{x}_{2i} + \cdots + \beta_k \bar{x}_{ki} + \alpha_i + \bar{\varepsilon}_i \\ (y_{it} - \bar{y}_i) &= \beta_2 (x_{2it} - \bar{x}_{2i}) + \cdots + \beta_k (x_{kit} - \bar{x}_{ki}) + (\varepsilon_{it} - \bar{\varepsilon}_i) \end{aligned}$$

- Consistent estimates of $\beta_2, ..., \beta_k$ are obtained as the part of the error (α_i) that was correlated with regressors has dropped
  - this relies on the nontestable assumption that only time-invariant part of the error is correlated with regressors.
- The fixed effects estimator can also be obtained by OLS estimation of y_{it} on a set of indicator variables for each individual and on $x_{2it}, ..., x_{kit}$.

Example: Wins and revenue for NBA teams

- What is the causal effect of winning last season on revenue this season
  - where need to control for some teams always having high revenue and winning more.
- Some NBA teams are on average always better than others.
- Dataset NBA has annual data on 29 teams for the 10 seasons 2001-02 to 2010-11
  - view as short panel dataset (T fixed and n large).

Data summary

- Variable description and summary statistics

Variable name	Storage type	Display format	Value label	Variable label
revenue	float	%8.0g		Team revenue in millions of 1999 $
lnrevenue	float	%9.0g		Natural logarithm of team revenue
wins	byte	%8.0g		Wins in previous season including playoff game
season	byte	%8.0g		1 to 10
teamid	byte	%8.0g		1 to 29

```
. summarize revenue lnrevenue wins season teamid
```

Variable	Obs	Mean	Std. dev.	Min	Max
revenue	286	95.71404	24.44207	58.49582	187.7212
lnrevenue	286	4.532293	.2359855	4.068955	5.234958
wins	286	41.03497	12.43758	9	67
season	286	5.541958	2.872126	1	10
teamid	286	14.86014	8.354935	1	29

Results

- Fixed effects estimate is that an extra win is associated with a 0.452% increase in revenue.

```
. * Pooled OLS
. regress lnrev wins season, vce(cluster teamid) noheader
                                (Std. err. adjusted for 29 clusters in teamid)
------------------------------------------------------------------------------
             |               Robust
   lnrevenue | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        wins |   .0068132   .0018965     3.59   0.001     .0029284    .0106981
      season |   .0182414   .0033079     5.51   0.000     .0114655    .0250172
       _cons |   4.151619   .0965789    42.99   0.000     3.953786    4.349452
------------------------------------------------------------------------------

. * Fixed effects
. regress lnrev wins season i.teamid, vce(cluster teamid) noheader
                                (Std. err. adjusted for 29 clusters in teamid)
------------------------------------------------------------------------------
             |               Robust
   lnrevenue | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        wins |   .0045238   .0008423     5.37   0.000     .0027985    .0062492
      season |   .0190451   .0035645     5.34   0.000     .0117436    .0263467
```

Further Details for grouped data

- The fixed effects estimator can also be applied to grouped data with model observations per group.
- For example, suppose we have data on many villages and on several individuals within each village.
- Then for individual i in group g

$$y_{gi} = \beta_1 + \beta_2 x_{2gi} + \cdots + \beta_k x_{kgi} + \alpha_g + \varepsilon_{gi}$$

- The fixed effects model assumes $u_{gi} = \alpha_g + \varepsilon_{gi}$ where only α_g is correlated with regressors.
- The cluster-specific fixed effects estimator is OLS in

$$y_{gi} - \bar{y}_g = \beta_2 (x_{2gi} - \bar{x}_{2g}) + \cdots + \beta_k (x_{kgi} - \bar{x}_{kg}) + \text{ error}$$

- Here $\bar{y}_g$, $\bar{x}_{2g}$,...,$\bar{x}_{kg}$ are averages across individuals in group g.
- Standard errors should cluster on group.
- Only parameters of regressors that vary within group can be estimated, as otherwise e.g. $x_{2gi} - \bar{x}_{2g} = 0$.

Further Details for panel data

- Often fixed effects estimates are much less precise as only within variation is used.
- For the fixed effects estimator, parameters of regressors that do not vary over time for a given individual cannot be estimated, since then $x_{it} = \bar{x}_i$ so $x_{it} - \bar{x}_i = 0$ for all $t = 1, ..., T$.
- An alternative (noncausal) estimator is the random effects estimator.
- The pooled model sets the model parameters to be the same for each individual and each time period
  - richer models can relax this.
- This model is a static model
  - richer dynamic models can be estimated such as $y_{it} = \beta_1 + \beta_2 y_{i,t-1} + \cdots$
  - but then the standard fixed effects estimator is inconsistent.
- In macroeconomics it is common to have n small while $T \longrightarrow \infty$.

References for panel data

- Basic fixed and random effects estimation is presented in many texts.

- A. Colin Cameron (2020), Analysis of Economics Data: An Introduction to Econometrics, chapter 17.2-17.3. https://cameron.econ.ucdavis.edu/aed/.
- Cunningham, Scott (2021), Causal Inference: The MixTape, Yale UP, chapter 8.
- A. Colin Cameron and Pravin K. Trivedi (2022), Microeconometrics using Stata: Volume 2, Second Edition, Stata Press, chapter 8.
- Joshua D. Angrist and Jörn-Steffen Pischke (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press, chapter 5.
- A. Colin Cameron and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications, Cambridge University Press, chapter 21.
- Jeffrey M. Wooldridge, (2010), Econometric Analysis of Cross Section and Panel Data, Second Edition, MIT Press, chapter 10.
- Andrew Gelman, Jennifer Hill and Aki Vehtari (2022), Regression and Other Stories, Cambridge University Press, chapter 21.1-21.2.