Advanced Quantitative Methods

Panel Data

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Today

- A shorter lecture on panel data & fixed effects
- Exercise

Cross-sectional data

- One unit is observed at a single point in time
  - e.g. typical survey data
- Very common form of data
- Regression model will examine *between*-unit variation
  - i.e. variation between respondents, interest groups, municipalities, states, etc.
- But we will worry that our variable of interest (e.g. a treatment) will be confounded by a bunch of other variables
  - e.g. Does private school cause an increase in income, or are those who go to private school smarter, more ambitious, etc.
- Can a *type* of data help us with this?

Panel data

- One unit is observed at multiple points in time
  - A person is surveyed before, during, and after an election campaign (multiple survey "waves")
  - A country has a certain policy in one year, but changes it in the next
  - A constituency has different vote shares for a far-right party across election years
- All else equal, panel data are essentially always better than cross-sectional data
- Why? Because we can examine *within*-unit variation
  - i.e. We compare a unit to itself
  - This is going to let us automatically control for a lot of observed and unobserved characteristics

Example: Donald Trump's 2016 election victory

- Many claim the working class voted for Trump because they were left behind economically
- If true, Democrats might want to respond with policy proposals to help the working class
- But were economic reasons the cause of the working class vote for Trump?
- Mutz (2018) tests whether the problem was instead "status threat" (i.e. basically racial status/threat)

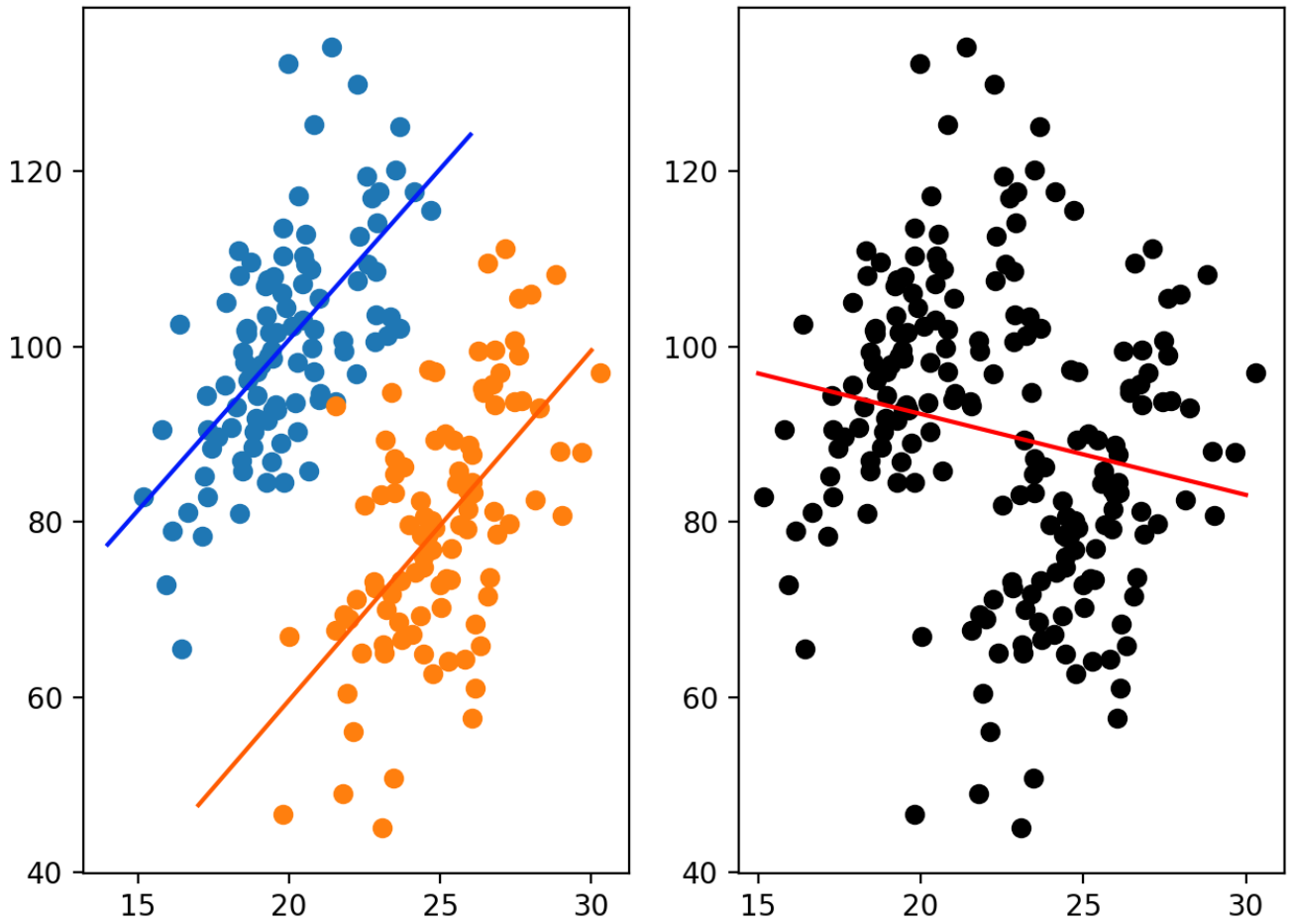
The problem

- But feelings of status threat is not assigned at random among US voters
- Any relationship to voting for Trump will be confounded by many individual-level characteristics
- Mutz (2018) thus takes advantage of a survey that asked the same individuals the same questions over time
  - That is, she has panel data
- She can thus exploit *within*-respondent variation
- This means she can account for the time-invariant characteristics of each respondent

Example data (from Mutz 2018)

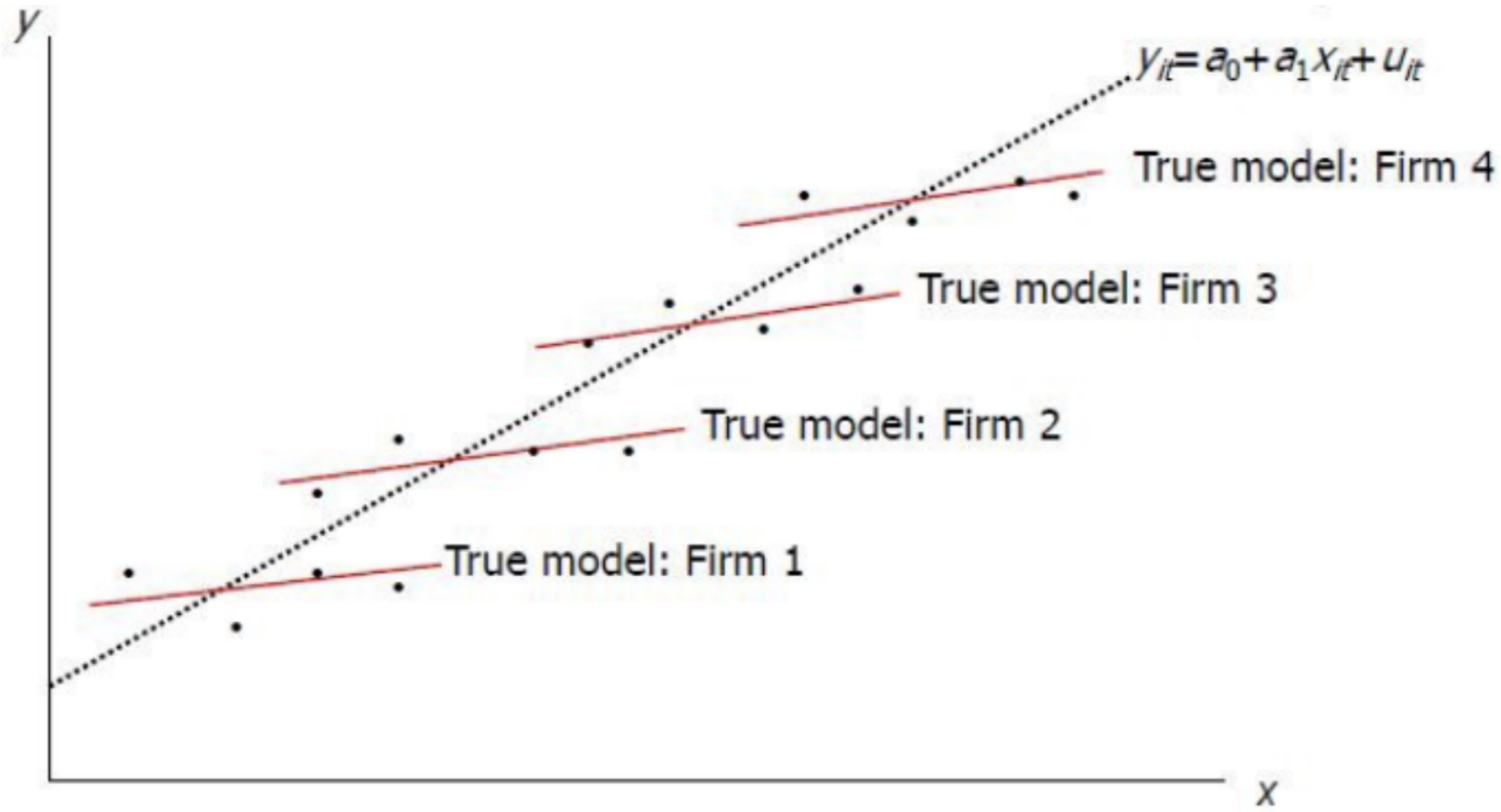
obs_number	id	cutdifftherm	sdo	income
1	6134	14	5	$100,000 to $124,999
2	6134	19	5.75	$75,000 to $84,999
3	7526	4	3.25	$30,000 to $34,999
4	7526	10	2	$25,000 to $29,999
5	7696	4	6.25	$60,000 to $74,999
6	7696	20	7.75	$50,000 to $59,999
7	20328	14	3.5	$35,000 to $39,999
8	20328	18	3.25	$60,000 to $74,999
9	26080	5	1.5	$50,000 to $59,999
10	26080	1	2.5	$75,000 to $84,999

Using within-unit variation



Left: separate intercept per group, **Right**: single global intercept

Each unit has its own intercept



The basic model setup

Cross-sectional data:

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{1}$$

Panel data (with fixed effects):

$$y_{it} = \alpha_i + \beta x_{it} + \epsilon_{it}, \tag{2}$$

y_{it} denotes the outcome for a unit i at time t (e.g. response from a specific person in a specific survey wave)

Importantly, α_i denotes a unit-specific intercept that captures factors for a unit (e.g. a person) that do not change over time

It eliminates factors that don't vary over time

The fixed effect α_i just represents a unit-level average

We can thus apply a "within" transformation to difference this out so that we don't need the parameter α_i at all

$$y_{it} - \bar{y}_i = (\alpha_i - \bar{\alpha}_i) + \beta(x_{it} - \bar{x}_i) + (\epsilon_{it} - \bar{\epsilon}_i) \tag{3}$$

- For every unit (e.g. respondent, district) the data are coded such the each outcome or variable is calculated as a deviation from its mean.
- e.g. If my status threat level increases, does my preference for Republicans increase or decrease relative to my average preference across time
- **Example R code to show this within transformation**

Can also take a "first difference"

$$y_{i,t} - y_{i,t-1} = (\alpha_i - \alpha_i) + \beta(x_{i,t} - x_{i,t-1}) + (\epsilon_{i,t} - \epsilon_{i,t-1}) \quad (4)$$

- More robust to temporal autocorrelation
- More robust to heteroskedasticity
  - Fixed effects: deviations from the average for the whole period.
  - First difference: deviation from last time period
- Fixed effects more efficient if no serial correlation (value of $y_{i,t}$ is uncorrelated with value of $y_{i,t-1}$)
- In practice, in political science we almost always use fixed effects

Benefits of fixed effects

- Factors that are unchanging over time do not drive our results
  - Personality
  - Gender
  - Race
  - Education (usually, unless it changes)
- This is a much stronger design than with cross-sectional data
  - Data that uses within-unit variation to model a relationship has fewer threats from confounding
- It is not a silver bullet, however
  - Variables that vary over time still are potential confounders
  - e.g. income, municipality revenue, immigration inflows, political attitudes
- Nevertheless, panel data with fixed effects substantially increase the credibility of a research design, and are thus much more credible than cross-sectional ones

Need to cluster standard errors

- With panel data, observations will be correlated within units
  - e.g. How a respondent answers a question at one point in time says something about how they will answer the same question at another point in time
- Thus, we use cluster-robust standard errors (easy to do in R)
  - Classical standard errors will often be far too conservative
- If you have very few clusters (e.g. you have data on just 10 regions), you need other methods to calculate your standard errors (e.g. boot-strapping)

Panel data as a stepping stone

- Understanding panel data and fixed effects are the starting point for understanding difference-in-differences
- Why? Because differences-in-difference use *within*-unit variation in the timing of policy implementations or events

Exercise solutions

```
# The respondent fixed effect allows us to capture this within-respondent
# variation
model_fe <- feols(cutdiffthermm ~ dem +
                                  income +
                                  lookingforwork +
                                  personeco +
                                  tradeper +
                                  tradeself +
                                  proimmself +
                                  chinaself +
                                  tradediffdem +
                                  immdiffdem +
                                  chinadiffdem +
                                  tradediffrep +
                                  immdiffrep +
                                  chinadiffrep +
                                  sdo +
                                  economy
                  | id,             # Respondent fixed effects
                  cluster = ~ id, # Cluster standard errors on each respondent
                                    # because each respondent provides multiple
                                    # responses in different survey waves at
                                    # different points in time
                  data = M)
```