Lecture 10: Panel data methods

Nic Rivers

Advanced quantitative research methods, API6319
Fall 2019

Cross sectional data and policy analysis

- Most of the data that we have encountered so far in the course is **cross-sectional** data
- Each observation is an individual at a specific snapshot in time:
  - Respondents in a survey
  - Countries (guns/violence; fertility/income)
- One of our aims is policy analysis. With regression:

$$\text{outcome}_i = \beta_0 + \beta_1 \text{policy}_i + \beta_2 \text{controls}_i + \epsilon_i$$

- We would like to interpret β_1 as the causal effect of 'policy' on 'outcome'. Is this possible? Under what circumstances?

Cross sectional data and policy analysis

- Most of the data that we have encountered so far in the course is **cross-sectional** data
- Each observation is an individual at a specific snapshot in time:
  - Respondents in a survey
  - Countries (guns/violence; fertility/income)
- One of our aims is policy analysis. With regression:

$$\text{outcome}_i = \beta_0 + \beta_1 \text{policy}_i + \beta_2 \text{controls}_i + \epsilon_i$$

- We would like to interpret β_1 as the causal effect of 'policy' on 'outcome'. Is this possible? Under what circumstances?
  - If policies are not randomly assigned to individuals, then it is likely that other factors are correlated with policy. We can try to control for these, but we may not be able to control for all determinants of policy implementation. In this case, the regression might suffer from omitted variables bias. We would not be able to infer the causal effect of the policy.

Example: beer taxes and traffic fatalities

- Roughly 40,000 highway traffic fatalities in US per year (over 2,000 in Canada)
- About one quarter involve a driver that was drinking
- One study estimates that one quarter of all drivers on road between 1 AM and 3 AM have been drinking (Levitt and Porter, 2001)
- There is interest in policies that might reduce traffic fatalities from impaired driving
- We will look at whether and how changes in the rate of taxation on beer affect traffic fatalities in the US
- Data on traffic fatalities by state as well as beer taxation by state every year from 1982 to 1988

Beer taxes and traffic fatalities

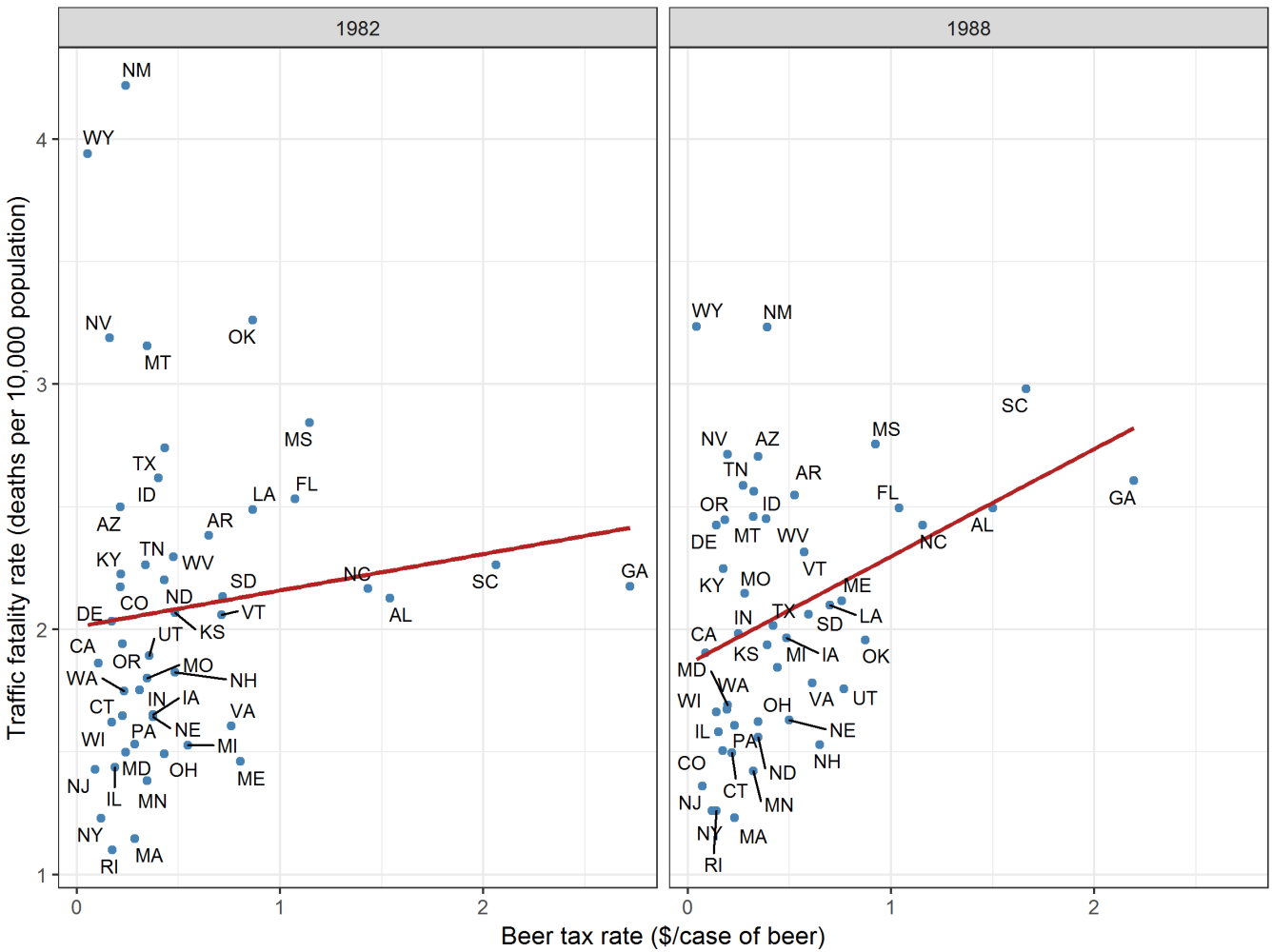
Estimated regression:

$$\text{fatalityrate} = \beta_0 + \beta_1 \text{beertax} + \text{other factors}$$

Beer taxes and traffic fatalities

Estimated regression:

$$\text{fatalityrate} = \beta_0 + \beta_1 \text{beertax} + \text{other factors}$$



Regression results

Table: Cross-sectional regression results

	Dependent variable:		
	fatrate		
	(1)	(2)	(3)
beertax	0.149	0.439**	0.365***
	(0.188)	(0.164)	(0.062)
Constant	2.010***	1.859***	1.853***
	(0.139)	(0.106)	(0.044)
Observations	48	48	336
R^2	0.013	0.134	0.093
Note:		*p<0.1; **p<0.05; ***p<0.01	

What should we take from these results?

- What do these regression results imply?
- What could be the problem?

What should we take from these results?

- What do these regression results imply?
- What could be the problem?
  - There could be omitted variables that are contaminating our estimates of the effect of beertax on fatalities:

What should we take from these results?

- What do these regression results imply?
- What could be the problem?
  - There could be omitted variables that are contaminating our estimates of the effect of beertax on fatalities:
    - Quality of vehicles driven in the state
    - Quality of highways
    - Rural or urban
    - Density of cars on the road
    - Social norms

What should we take from these results?

- What do these regression results imply?
- What could be the problem?
  - There could be omitted variables that are contaminating our estimates of the effect of beertax on fatalities:
    - Quality of vehicles driven in the state
    - Quality of highways
    - Rural or urban
    - Density of cars on the road
    - Social norms
  - How should we fix this potential omitted variable problem?

What should we take from these results?

- What do these regression results imply?
- What could be the problem?
  - There could be omitted variables that are contaminating our estimates of the effect of beertax on fatalities:
    - Quality of vehicles driven in the state
    - Quality of highways
    - Rural or urban
    - Density of cars on the road
    - Social norms
  - How should we fix this potential omitted variable problem?
    - One approach: collect data on all of these factors and include in regression (could be difficult)
    - Another approach: leverage panel data

Before/after comparison

- Imagine that there is an omitted variable that varies between states that we don't measure, such as:
  - Social norms towards drinking/driving
  - Highway quality
- Let's call this variable Z_i. The population regression line is:

$$\text{fatalityrate}_{it} = \beta_0 + \beta_1 \text{beertax}_{it} + \beta_2 Z_i + u_{it}$$

- Note it subscript - we observe states (i) over time (t)
- Our estimates for β_1 might be biased, since they did not include Z_i

Before/after comparison

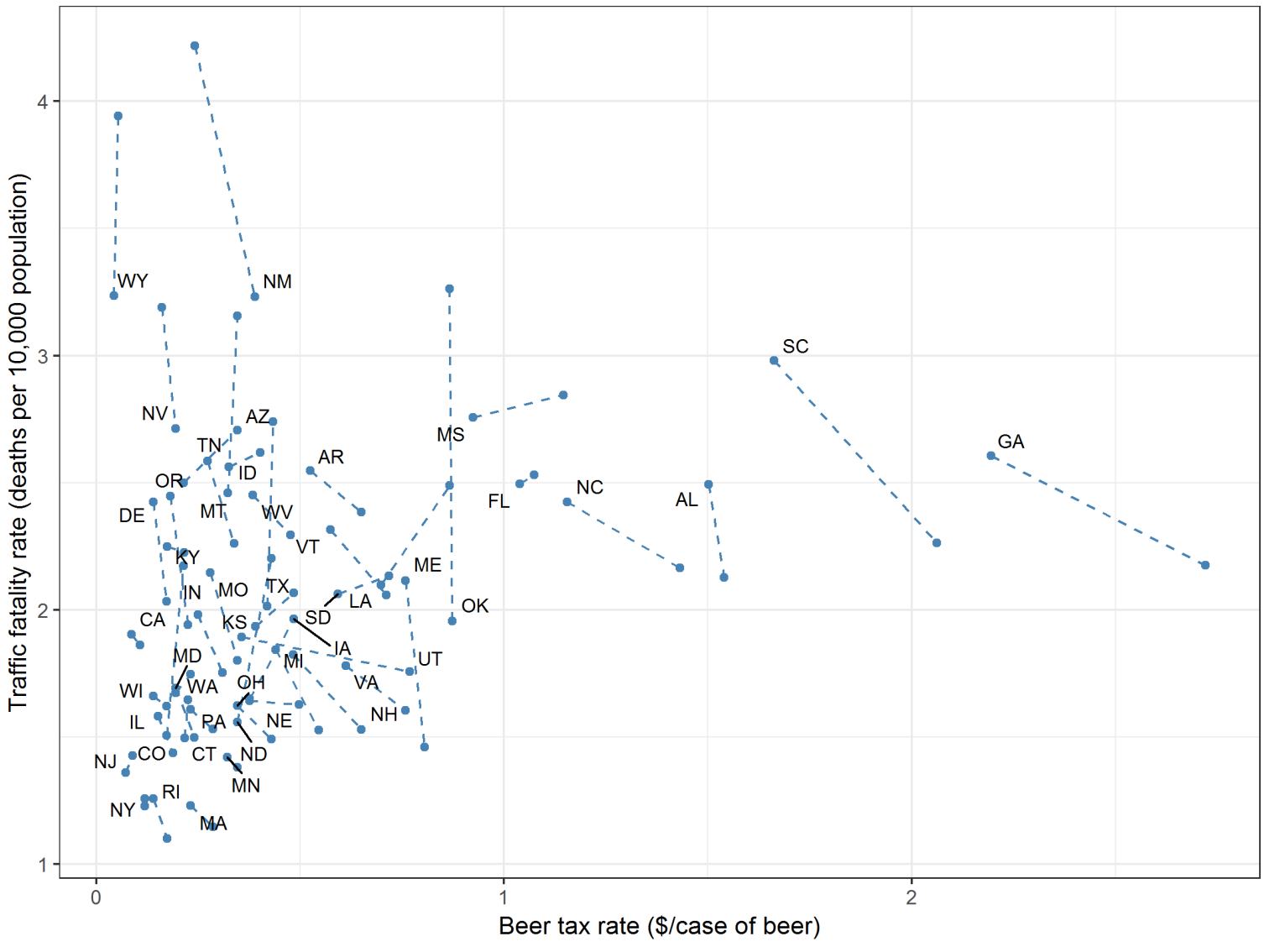
- Our estimates using 1982 and 1988 data both suffer from omitted variable bias

$$\text{fatalityrate}_{i1982} = \beta_0 + \beta_1 \text{beertax}_{i1982} + \beta_2 Z_i + u_{i1982}$$
$$\text{fatalityrate}_{i1988} = \beta_0 + \beta_1 \text{beertax}_{i1988} + \beta_2 Z_i + u_{i1988}$$

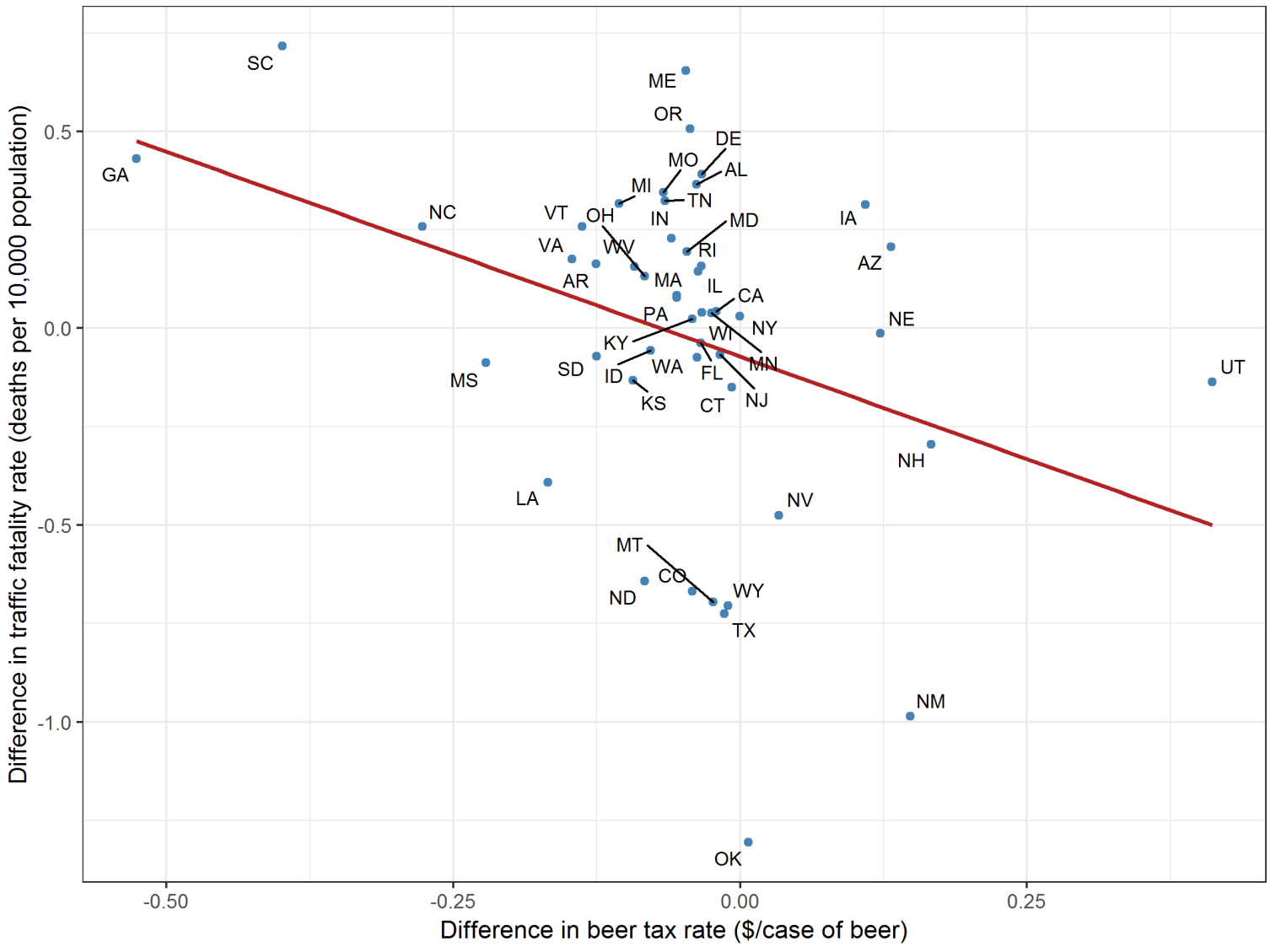
- But, if we focus on differences over time, we can eliminate the effect of Z_i:

$$\begin{aligned} &\text{fatalityrate}_{i1988} - \text{fatalityrate}_{i1982} \\ &= \beta_1 \left(\text{beertax}_{i1988} - \text{beertax}_{i1982}\right) + u_{i1988} - u_{i1982} \end{aligned}$$

Differences



Differences (2)



Regression results

Table: Cross-sectional vs. difference-in-difference regression results

	Dependent variable:	
	fatrate 1982 and 1988 data	diff_fatrate 1982 to 1988 change
	(1)	(2)
beertax	0.268^{**}	
	(0.126)	
diff_tax		-1.040^{**}
		(0.417)
Constant	1.944^{***}	-0.072
	(0.087)	(0.061)
Observations	96	48
R^2	0.046	0.119
Note:		*p<0.1; **p<0.05; ***p<0.01

Interpretation of results

- A one dollar per case increase in beer tax is associated with a reduction in traffic fatalities of 1.041 per 10,000 people.
- Average amount of traffic fatalities is 2/10,000 people.
- This is potentially a very big effect
- It is statistically significant as well
- Probably a much more credible estimate of the effect of beertax on fatalities, since it eliminates fixed factors in each state that could be correlated with beertax and fatalities

Panel data

- Assume the population regression function is:

$$\text{outcome}_{it} = \beta_0 + \beta_1 \text{treatment}_{it} + \beta_2 Z_i + e_{it}$$

- Z_i is some time-invariant feature of individuals in the data set that could be correlated with both outcome and treatment. We can't easily measure it.
- One approach to estimating β_1 in this case is to look at changes, rather than levels
- By looking at changes, we eliminate the effect of Z_i, since it remains constant:

$$\text{outcome}_{i2} - \text{outcome}_{i1} = \beta_1(\text{treatment}_{i2} - \text{treatment}_{i1}) + u_{it}$$

- This approach is known as **differencing** (sometimes known as a **difference-in-difference** approach)

Fixed effects

- An alternative to differencing is to estimate the regression model with **fixed effects**
- Fixed effects are dummy variables for each individual in the data set, which capture the effect of Z_i. Now we have:

$$\text{outcome}_{it} = \beta_0 + \beta_1 \text{treatment}_{it} + \alpha_i + e_{it}$$

  where α_i is a dummy variable for each state/unit/individual in the data set. These can be unobserved things like 'attitude' or 'culture' (as well as potentially-observable things)
- α_i captures individual-specific factors that do not vary over time
- We can now include data from a number of different years, rather than just looking at changes between a pair of years as we did in the previous analysis

Other variables

- Fixed effects control for any factors that are constant over time but differ between individuals
- This can eliminate a major portion of omitted variable bias
- However, it may not solve all problems of omitted variable bias
- There could be variables that vary over time that are correlated with treatment variable and also affect outcome variable
- As with other multiple regression analysis, we can try to control for some of these

Time fixed effects

- If there are factors that affect all of the individuals in the sample in a similar way, and these vary over time, then we could be worried about omitted variable bias from not including these
- In the traffic fatalities example, some of these factors might be:

Time fixed effects

- If there are factors that affect all of the individuals in the sample in a similar way, and these vary over time, then we could be worried about omitted variable bias from not including these
- In the traffic fatalities example, some of these factors might be:
  - Changes in vehicle technology (e.g., airbags)
  - General changes in social norms regarding drinking and driving
  - Changes in federal laws

Time fixed effects

- If there are factors that affect all of the individuals in the sample in a similar way, and these vary over time, then we could be worried about omitted variable bias from not including these
- In the traffic fatalities example, some of these factors might be:
  - Changes in vehicle technology (e.g., airbags)
  - General changes in social norms regarding drinking and driving
  - Changes in federal laws
- Some of these are measurable; some are difficult to measure.
- We can include both using dummy variables for each year

$$\text{outcome}_{it} = \beta_0 + \beta_1 \text{treatment}_{it} + \alpha_i + \lambda_t + e_{it}$$

- λ_t is a dummy variable for each year in the data
- How would we interpret β_1 now?

Time fixed effects

- If there are factors that affect all of the individuals in the sample in a similar way, and these vary over time, then we could be worried about omitted variable bias from not including these
- In the traffic fatalities example, some of these factors might be:
  - Changes in vehicle technology (e.g., airbags)
  - General changes in social norms regarding drinking and driving
  - Changes in federal laws
- Some of these are measurable; some are difficult to measure.
- We can include both using dummy variables for each year

$$\text{outcome}_{it} = \beta_0 + \beta_1 \text{treatment}_{it} + \alpha_i + \lambda_t + e_{it}$$

- λ_t is a dummy variable for each year in the data
- How would we interpret β_1 now?
  - Holding constant all individual-specific factors as well as all time-specific factors, outcome changes on average by β_1 units when treatment changes by one unit

Other variables

- By including individual fixed effects (dummy variables), we have controlled for all factors that vary over individuals, but do not vary over time
- By including time fixed effects (dummy variables), we have controlled for all factors that vary over time, but do not vary over individuals
- There may still be omitted variables that contaminate our estimates. How?

Other variables

- By including individual fixed effects (dummy variables), we have controlled for all factors that vary over individuals, but do not vary over time
- By including time fixed effects (dummy variables), we have controlled for all factors that vary over time, but do not vary over individuals
- There may still be omitted variables that contaminate our estimates. How?
- Variables that vary over time and that are specific to an individual, and which affect both the dependent and independent variables
- We can try to control for these in the way that we normally do in multiple regression:

$$\text{outcome}_{it} = \beta_0 + \beta_1 \text{treatment}_{it} + \beta_2 X_{it} + \beta_3 Y_{it} + \cdots + \alpha_i + \lambda_t + e_{it}$$

where X and Y are other independent variables

Regression results

Table: Fixed effects regression results

	Dependent variable:			
	fatrate			
	OLS		*felm*	
	OLS	F.E.	F.E.+T.E.	F.E.+T.E.+controls
	(1)	(2)	(3)	(4)
beertax	0.365^{***}	-0.655^{**}	-0.639^{*}	-0.545
	(0.062)	(0.315)	(0.386)	(0.359)
unrate				-0.094^{***}
				(0.016)
yngdrv				0.864
				(0.929)
dry				0.020^{*}
				(0.012)
Constant	1.853^{***}			
	(0.044)			
Observations	336	336	336	336
R^2	0.093	0.905	0.909	0.934
Note:				*p<0.1; **p<0.05; ***p<0.01

Estimating in R

- You can estimate a panel model using lm() by just including the dummy variables that you require.
  lm(fatrate $\sim$ beertax + factor(state) + factor(year), data=dat)
- There are a lot of these (48 states + 7 years = 55 fixed effects). There aren't normally of much interest. We typically don't report them.
- The lfe package is designed for panel data. Use the felm() fuction, which is an extension of the lm() function that you already know.
  `felm(fatrate ~ beertax | factor(state) + factor(year), data=dat)`
- To account for non-random sampling:
  `felm(fatrate ~ beertax | factor(state) + factor(year) | 0 | state, data=dat)`

Panel data advantages

- Significantly increase the size of data set (increase precision of estimates)
- Before/after comparison of effect of public policy intervention is compelling
- Ability to control for unobserved (and observed) effects that are both time-invariant but differ by individual as well as effects that are individual-invariant but differ over time
- This gets us closer to a causal estimate of treatment. Still need to be cautious (e.g., choice of policy could be driven by outcome (endogenous))