Traffic Death Analysis

In this exercise, we will be analyzing the effect of alcohol taxes on traffic death in the United States. The data set used in this exercise, `fatalities.csv`, is a state-year panel dataset (meaning it includes data on multiple states, and the data includes several years of data for each state. The data contains 336 observations on 34 variables. The variables used in the exercise are defined as follows:

`state`: factor variable indicating states

`year`: factor variable indicating years

`beertax`: numeric variable, Tax on the case of beer

In these exercises, we'll be looking at how beer taxes (which are believed to reduce alcohol consumption, potentially reducing drunk driving deaths) impact car accident fatality rates.

More specifically, though, we'll be approaching our estimation of the impact of beer taxes in a few different ways in an effort to give you more of an intuitive sense of what happens when you add fixed effects to a regression.

Exercise 1

Download and load the data from this link [https://github.com/nickeubank/MIDS_Data/blob/master/UDS_arrest_data.csv], or by going to http://www.github.com/nickeubank/MIDS_Data/ [http://www.github.com/nickeubank/MIDS_Data/] and downloading the `us_driving_fatalities.csv` dataset.

How many states does this dataset contain? What's the time frame of this dataset? (From which year to which year). And what constitutes a single observation (i.e. what is the unit of analysis for each row of the data?)

Exercise 2

We use the fatality rate per 10,000 as the dependent variable. Construct this variable. Name it as `fat_rate`. Hint: You can compute it using total fatalities(`fatal`) and population (`pop`). Note that because `pop` is often the name of a method in Python, you may have to navigate around some issues.

Exercise 3

Draw a scatter plot using `beertax` as the x-axis, and `fat_rate` as the y-axis. Draw a fitted line showing the correlation between these two variables

Exercise 4

Fit a simple OLS regression. This is what is called a "pooled" regression because we're "pooling" observations from different years into a single regression. What do your results imply about the relationship between Beer Taxes and fatalities?

$$FatalityRate_i = \beta_0 + \beta_1 \times BeerTax_i$$

Exercise 5

Now estimate your model again, this time adding state fixed effects (using the `C()` notation and your normal linear model machinery). What does this result imply about the relationship between beer taxes and fatalities?

Exercise 6

Explain whey your results in Exercises 4 (without fixed effects) and Exercise 5 (with state fixed effects) look so different. What does this imply about states with high beer taxes?

Fixed Effects by Demeaning

Rather than just add indicator variables, we'll now use a different strategy for estimating fixed effects called an "entity-demeaning." This method is more computationally efficient, and can also help you understand how fixed effects work.

Let's begin by assuming we want to estimate the following fixed-effect model:

$$FatalityRate_{it} = \alpha + \beta BeerTax_{it} + \Psi Z_i + \epsilon_{it} \qquad (1)$$

Where $FatalityRate_{it}$ is the fatality rate of state i in year t, $\beta BeerTax_{it}$ is the beer tax of state i in year t. Z_i is a state fixed effect.

Rather than adding indicator variables, however, we'll use entity-demean as follows:

First, we take the average on both sides of the regression. Here n is the number of periods.

$$\frac{1}{n}\sum_{i=1}^{n} FatalityRate_{it} = \alpha + \beta_1 \frac{1}{n}\sum_{i=1}^{n} BeerTax_{it} + \Psi \frac{1}{n}\sum_{i=1}^{n} Z_i + \frac{1}{n}\sum_{i=}^{n}$$

$$\overline{FatalityRate}_i = \alpha + \beta_1 \overline{BeerTax}_i + \Psi Z_i + \overline{\epsilon}_i.$$

Substracting the from the main equation yields:

$$FatalityRate_{it} - \overline{FatalityRate}_i = \beta_1 (BeerTax_{it} - \overline{BeerTax}_i) + \Psi(Z_$$

$$\widetilde{FatalityRate}_{it} = \beta_1 \widetilde{BeerTax}_{it} + \widetilde{\epsilon}_{it}$$

Where the $\sim$ means values have been demeaned by group.

By taking the difference between the value of each observation (state-year) and the mean value of the entity (state) over n periods, we analyze how the within-state variation of beer tax affects that of the fatality rate. Moreover, by doing so we no longer need to estimate the fixed effects of Z_i, saving computing power if we are working on a dataset with a large number of fixed effects.

Exercise 6

Implement the above entity-demeaned approach to estimate the fixed-effects model by hand (use basic functions, not full tools like `PanelOLS` or `C()` notation in `python`, or `lfe` or `C()` notation in `R`).

Exercise 7

Fit the model with state fixed-effect using `PanelOLS` / `lfe`. Compare it to your by-hand output. Interpret the result.

Exercise 8

Now (using `PanelOLS` or `lfe`) estimate a fixed effects model using the following specification. Add fixed effects for **both** the state and the year, as well as the other covariates you think are important X_{it}).

(Note: you may want to make sure `PanelOLS` adds an intercept to aid in interpretation of controls if you include categorical variables. If you use `PanelOLS.from_formula()`, just put a `1+` in after the ~, or add a column of 1s to your X matrix if you're working with numpy arrays.)

Explain (a) the type of phenomenon we control for by adding `year` fixed effects, and

(b) your choice of covariates. Cluster the standard error at the state level. Interpret the result.

$$FatalityRate_{it} = \alpha + \beta BeerTax_{it} + X_{it} + State_i + Year_t + \epsilon_{it}$$

Absolutely positively need the solutions?

Don't use this link until you've really, really spent time struggling with your code! Doing so only results in you cheating yourself.

Link [../solutions_warning.html]