Empirical Economics - Data exercise II

Information:

The deadline for this exercise is **3 December, 23:59**. The assignment can be done in groups of three students. For this lab assignment, you should paste the *relevant* output in the file where you include your answers. Always also include your R-code so that we can check what you did (and possibly see what went wrong).

Introduction:

Since July 2019 you are not allowed to use any electronic devices while driving in the Netherlands. This law was passed universally, making it difficult to empirically analyse the effects of this texting ban on the total number of road accidents. In the US, however, each state decides unilaterally on the imposition of texting bans. In this data-exercise you will make use of the dataset textingbans.csv to empirically analyse the effect of texting bans on road accidents.[1]

The dataset contains the following variables:

state	indicator for state
time	indicator for time
accident	total number of accidents in state at time t
year	year
pop	population of state at time t
rgastax	real prevailing gas tax in state at time t
unemp	unemployment rate in state at time t
txmsban	state had texting ban at time t (0/1)
treated	state introduced a texting ban at any point time (0/1)

Part 1: Panel data models

Use the dataset textingbans.csv that is provided on Canvas.

1. Perform a pooled regression to estimate the effect of a texting ban on the total number of accidents. Note: to estimate this regression, you

[1] This dataset is taken from the paper "Texting Bans and Fatal Accidents on Roadways: Do They Work? Or Do Drivers Just React to Announcements of Bans?" by Abouk and Adams (AEJ: Applied, 2013)

will first need to create the variable $log(accident)$ yourself.

$$\log(\text{accident})_{st} = \beta_0 + \beta_1 \cdot \text{txmsban}_{st} + \varepsilon_{st}.$$

Interpret the estimated coefficient β_1 (discuss both effect size and significance). What happens to β_1 if you estimate the same regression controlling for log population size? Why is it important to include population size here?

2. You next want to exploit the panel nature of the dataset and estimate a fixed effects model. Write down the model that you would estimate. Estimate the model and interpret the estimate for the effect of the textingban on log(accident).

3. Re-estimate you fixed effects model additionally controlling for log(unemp), log(permale), and log(rgastax). Pick one of these three control variables and explain in words what it potentially controls for.

4. So far, we have not yet controlled for time trends. Re-estimate you model including a linear time trend.

5. Generate a new variable that contains the mean number of accidents (over time) for all states together. Plot the mean number of accidents over time. Do you think the linear trend in the previous question was the right way to control for time? Explain your answer (include the graph in your answer).

6. Think of a better way to control for time, and use this to re-estimate your model from question 4.

7. Do you think you can interpret the effect of the texting ban on accidents from your last model as *causal*? Motivate your answer.

Difference in Differences

Using the same dataset you can also perform a difference-in-difference analysis. Keep two states of your choice, but make sure that you pick one state that never introduced a texting ban (control state), and one state that introduced a texting ban at some point during the observation period (treatment state). Ideally, make sure that the two states do not differ too much in terms of population size.

8. Provide descriptives of the two states you picked. Do the states look similar in terms of population size, percentage males, unemployment rates and the gasoline tax? Based on the summary statistics, argue whether you expect a difference-in-difference analysis that uses these two states to yield valid estimates (i.e. estimates that can be interpreted causally).

9. Create a graph that shows the log number of accidents over time for both states separately. Include a vertical line in your graph that indicates the time of the introduction of the texting ban in the "treatment" state. Based on your graph, did the two states have parallel trends in the number of accidents leading up to the texting ban in the treatment state? Include your graph in the answer.

10. Create a dummy *treat_state* that is equal to one if the state is your treatment state, a dummy *post* that is equal to 0 before the introduction of the texting ban and equal to 1 after. Estimate the following difference-in-differences model:

$$\log(\text{accident})_{st} = \beta_0 + \beta_1 \text{treat\_state}_{st} + \beta_2 \text{post}_{st} + \beta_3 \text{treat\_state}_{st} * \text{post}_{st} + \varepsilon_{st}.$$

Give a precise interpretation of all your estimated coefficients.

Part 3: Binary choice models

The decision to introduce a texting ban is made at the state level, but is likely not random. This is what we will investigate in the remainder of the assignment. For this part of the assignment, keep all states in the dataset but only use observations for which $time = 1$.

11. Describe which state characteristics are correlated with the introduction (at any point in time) of a texting ban. To do so, estimate the following model using a linear probability model (OLS). Make sure to use the right standard errors.

$$\text{treated}_s = \beta_0 + \beta_1 \cdot \log(\text{population})_s + \beta_2 \cdot \log(\text{accident})_s + \varepsilon_s$$

Give a precise interpretation of your estimated coefficients β_1 and β_2. Is this what you expected? Can you think of an explanation for what you find?

12. Generate the predicted values for the first model and discuss them. Next, use Probit and Logit to estimate the same model specification as in question (10). Do your coefficient estimates confirm the results of the linear probability model? Calculate the marginal effect of each coefficient at the mean, and interpret the result.

13. Compute the average marginal effect of each coefficient from both the Logit and the Probit regression. Compare these marginal effects to the corresponding estimated marginal effects in (2).

14. Which model do you prefer for estimating the model from question 10 and why? And can you include state fixed effects in this model? Why/why not?

15. Include you R-code as an appendix to this assignment.