EC2227 Labs, Problem Set 1

Due date: Sunday, March 21, 11.59 pm EST

Instructions

- This homework consists of 2 questions, for a total of 100 points.
- Create a do–file with your code, an automated log–file of your answers from that code, and fill out the provided Word template with your final answers as indicated.
- Submit all three files (i.e. do–file, log–file, and Word document) online via the Canvas homework tool. Failure to submit 1 out of 3 files will determine a 33.33% penalty in this homework grade. Failure to submit 2 out of 3 files will determine a 66.66% penalty in this homework grade.
- You are allowed to work with one partner, but each student needs to submit solutions on Canvas. If you do work with a partner, you should **clearly state your partner's name** in your submission.
- For any questions about this assignment, please contact your instructor. For any questions about your grade, please contact Haydar Evren.

Question 1 (45 points): Load the `discrim` data set using the `bcuse` command. This data set contains prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. Read all the variable labels carefully.

For purposes of this question, use variables from the first wave as indicated by the labels whenever two waves for the same variable are available. For instance use `wagest` and not `wagest2` unless explicitly asked.

1. What is the structure of the data? (Cross-section, Time Series, or Panel data) (2 pt)
2. How many fast-food restaurants are there in the data? (2 pt)
3. How many of the restaurants are in New Jersey? (2 pts)
4. What is the average starting wage across all restaurants? (2 pts)
5. Define what a dummy variable is and name all the dummy variables in this dataset. (2+2=4 pts)
6. According to data available on the FRED website, the minimum wage in New Jersey and Pennsylvania in 1995 was $4.25. Generate a new dummy variable called `above_min_wage` which takes value one for restaurants where the starting wage is above $4.25 and 0 otherwise. (5 pts)
7. What is the average starting wage in restaurants in New Jersey and Pennsylvania, respectively? (Hint: Combine `bysort` and `summarize`.) (5 pts)
8. Do all the fast-food restaurants in this data belong to a chain? (3 pt)

9. Following the variable label of the variable `chain`, label values for this variable. (Hint: You will need to first define values and create a value label using the `label define` command. Then you will need to use `label value` in order to assign this label to the variable.) (4 pts)

10. Now relabel the the variable `chain` as "Chain that the restaurant belongs to". (3 pt)

11. Create a new variable called `avg_price` that takes the average price of fries for the chain that the restaurant belongs to. (Hint: Combine `bysort` and `egen`.) (5 pts)

12. One of your lifestyle blogger friends just glanced over your shoulder and saw what you are working on. Now they are pestering you to give them average prices for fries, coke and entres of all four chains in an excel file (your friend is not an Econ major). To accomplish this you will need to use the `collapse`. To know more about this command, use `help collapse` in Stata. Once you have successfully created the desired data set, use `export excel` to save it by the name 'averages.xlsx'. (4+4 = 8 pts)

    The point of the problem is to enable you to be able to use commands you haven't seen before by learning to use Stata help.

Question 2 (55 points):

Your overhear your friend quoting an Instagram post which claimed that the prison population is actually lower in states with higher crime rates. Ever vigilant against fake news, you decide to test this claim for yourself.

Download and save the datasets `prison` and `murder` to your preferred working folder created for the class.

1. Open the `prison` file [Hint: you need to use the command `use`]. Is this a cross-section, time series, or panel dataset? (2 pts)

2. Keep only the observations for the 1990s (`year >= 90`) [Hint: use `keep if`]. Save this new 1990s-only dataset as `prison_90s`. Is this new dataset a cross-section, time series, or panel dataset? How many observations are there? (5 pts)

3. Generate a correlation matrix (`corr`) for the four variables `unem`, `criv`, `crip` and `pris`. Which of the former three variables has the strongest correlation with `pris`, and is it positive or negative? Does this support your friend's claim? (5 pts)

4. Produce a `twoway` scatterplot of `pris` over `criv`. Make sure you include a line of best fit. Is the relationship between the two variables positive or negative? (Note: you do not need to paste your graph into your Word document.)(5 pts)

    You talk to your friend. She appreciates your findings, but adds that you did not quite understand the Instagram post properly. Specifically, it claimed that putting more people in prison acts as a deterrent, so that crime decreases in the following year.

5. Reopen `prison`, and now keep only the observations for Massachusetts (`state` == 22). Save this MA-only dataset as `prison_MA`. Is this a cross-section, time series, or panel dataset? (3 pts)

6. Generate `pris_gro` equal to the growth rate of the prison population. The growth rate of any variable is given by the following equation: (8 pts)

$$\text{Growth Rate}(x_t) = \frac{x_t - x_{t-1}}{x_{t-1}} \times 100$$

   In order to generate `pris_gro`, you will need to:

   - Sort your dataset by year [Hint: use `sort year`].
   - Use `pris[_n-1]` for x_{t-1} in the formula above.

   What is the average of the new variable `pris_gro`?

7. Your friend's claim is that an increase in prison population causes crime to decrease in the *following* year, not in the same year. Therefore, generate the lagged growth rate, `pris_gro_lag`, using the following syntax: `generate pris_gro_lag = pris_gro[_n-1]`. (2 pts)

8. Find the correlation coefficient between `criv` and `pris_gro_lag`. Create a scatterplot of the two variables, and include a line of best fit. Is your friend correct in the case of Massachusetts? (5 pts)

9. Reopen `prison`, and repeat 6-7 for all states simultaneously. You will need to do the following: (10 pts)

   - Sort your dataset by state and year [Hint: use `sort state year`]
   - In order to generate `pris_gro` for all states simultaneously, you will need to type `by state:` in front of the `generate` command.
   - In order to generate `pris_gro_lag` for all states simultaneously, you will need to type `by state:` in front of the `generate` command.

10. Fascinated, your friend rereads the Instagram post. It actually referred to murder only, not to crime in general. Merge the dataset `murder` into your current (`prison`) dataset. In order to do so, you will need to do the following: (10 pts)

    - Inspect `murder` to verify that the variables `state` and `year` go by the same names there. (Check the following: Is the state variable in `murder` called `state`, or is it called something different? Is the year variable in `murder` called `year`, or is it called something different?)
    - If the variable names do not match, change the current names in `prison` to match those in `murder`.
    - Use the `merge` command to merge the two datasets [Hint: this is a one-to-one merge].

    Evaluate the correlation between `mrdrte` and `pris_gro_lag`, and produce a scatterplot and line of best fit for these two variables.