EC2227 Labs, Problem Set 1

Haydar Evren, Dinghe Cui, Arnab Palit

Grading TA: Haydar Evren; evrenh@bc.edu

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)

The data is Panel Data as it has observations across different cross sections in time.

. describe

Contains data from E:\ONLINE WORKING\BioInformatics\_Stata\_Modelling\discrim - Copy.dta

obs: 410

vars: 37 8 Jan 2002 22:26

size: 47,150

--------------------------------------------------------------------------------------------------------

storage display value

variable name type format label variable label

--------------------------------------------------------------------------------------------------------

psoda float %9.0g price of medium soda, 1st wave

pfries float %9.0g price of small fries, 1st wave

pentree float %9.0g price entree (burger or chicken), 1st wave

wagest float %9.0g starting wage, 1st wave

nmgrs float %9.0g number of managers, 1st wave

nregs byte %9.0g number of registers, 1st wave

hrsopen float %9.0g hours open, 1st wave

emp float %9.0g number of employees, 1st wave

psoda2 float %9.0g price of medium soday, 2nd wave

pfries2 float %9.0g price of small fries, 2nd wave

pentree2 float %9.0g price entree, 2nd wave

wagest2 float %9.0g starting wage, 2nd wave

nmgrs2 float %9.0g number of managers, 2nd wave

nregs2 byte %9.0g number of registers, 2nd wave

hrsopen2 float %9.0g hours open, 2nd wave

emp2 float %9.0g number of employees, 2nd wave

compown byte %9.0g =1 if company owned

chain byte %9.0g BK = 1, KFC = 2, Roy Rogers = 3, Wendy's = 4

density float %9.0g population density, town

crmrte float %9.0g crime rate, town

state byte %9.0g NJ = 1, PA = 2

prpblck float %9.0g proportion black, zipcode

prppov float %9.0g proportion in poverty, zipcode

prpncar float %9.0g proportion no car, zipcode

hseval float %9.0g median housing value, zipcode

nstores byte %9.0g number of stores, zipcode

income float %9.0g median family income, zipcode

county byte %9.0g county label

lpsoda float %9.0g log(psoda)

lpfries float %9.0g log(pfries)

lhseval float %9.0g log(hseval)

lincome float %9.0g log(income)

ldensity float %9.0g log(density)

NJ byte %9.0g =1 for New Jersey

BK byte %9.0g =1 if Burger King

KFC byte %9.0g =1 if Kentucky Fried Chicken

RR byte %9.0g =1 if Roy Rogers

--------------------------------------------------------------------------------------------------------

Sorted by:

.

1. How many fast-food restaurants are there in the data? (2 pt)
2. How many of the restaurants are in New Jersey? (2 pts)

There are 331 restaurant in New Jersey

NJ=1

. codebook state

--------------------------------------------------------------------------------------------------------

state NJ = 1, PA = 2

--------------------------------------------------------------------------------------------------------

type: numeric (byte)

range: [1,2] units: 1

unique values: 2 missing .: 0/410

tabulation: Freq. Value

331 1

79 2

1. What is the average starting wage across all restaurants? (2 pts)

The means for the two waves are 4.615641 for wagest and 4.996273 for wagest2

. summarize wagest wagest2

Variable | Obs Mean Std. Dev. Min Max

-------------+---------------------------------------------------------

wagest | 390 4.615641 .3470151 4.25 5.75

wagest2 | 389 4.996273 .2531905 4.25 6.25

.

1. Define what a dummy variable is and name all the dummy variables in this dataset. (2+2=4 pts)

A dummy variable is a variable that takes on the values 1 and 0; 1 means something is true (such as age < 25, sex is male, or in the category “very much”). Dummy variables are also called indicator variables. A dummy variable is a variable that takes on the values 1 and 0; 1 means something is true (such as age < 25, sex is male, or in the category “very much”). Dummy variables are also called indicator variables.

The following are the dummy variables compown state NJ BK KFC RR

. codebook compown state NJ BK KFC RR

--------------------------------------------------------------------------------------------------------

compown =1 if company owned

--------------------------------------------------------------------------------------------------------

type: numeric (byte)

range: [0,1] units: 1

unique values: 2 missing .: 0/410

tabulation: Freq. Value

269 0

141 1

--------------------------------------------------------------------------------------------------------

state NJ = 1, PA = 2

--------------------------------------------------------------------------------------------------------

type: numeric (byte)

range: [1,2] units: 1

unique values: 2 missing .: 0/410

tabulation: Freq. Value

331 1

79 2

--------------------------------------------------------------------------------------------------------

NJ =1 for New Jersey

--------------------------------------------------------------------------------------------------------

type: numeric (byte)

range: [0,1] units: 1

unique values: 2 missing .: 0/410

tabulation: Freq. Value

79 0

331 1

--------------------------------------------------------------------------------------------------------

BK =1 if Burger King

--------------------------------------------------------------------------------------------------------

type: numeric (byte)

range: [0,1] units: 1

unique values: 2 missing .: 0/410

tabulation: Freq. Value

239 0

171 1

--------------------------------------------------------------------------------------------------------

KFC =1 if Kentucky Fried Chicken

--------------------------------------------------------------------------------------------------------

type: numeric (byte)

range: [0,1] units: 1

unique values: 2 missing .: 0/410

tabulation: Freq. Value

330 0

80 1

--------------------------------------------------------------------------------------------------------

RR =1 if Roy Rogers

--------------------------------------------------------------------------------------------------------

type: numeric (byte)

range: [0,1] units: 1

unique values: 2 missing .: 0/410

tabulation: Freq. Value

311 0

99 1

.

1. 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)
2. What is the average starting wage in restaurants in New Jersey and Pennsylvania, respectively? (Hint: Combine bysort and summarize.) (5 pts)
3. Do all the fast-food restaurants in this data belong to a chain? (3 pt)
4. 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 labeldefine command. Then you will need to use labelvalue in order to assign this label to the variable.) (4 pts)
5. Now relabel the the variable chain as “Chain that the restaurant belongs to”. (3 pt)
6. 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)
7. 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):

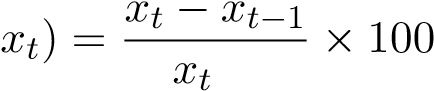
Your overhear your friend quoting an Instagram post which claimed that the prison population is actually lower in states with higher crime rates. Ever vigilant against fake news, you decide to test this claim for yourself.

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.

1. 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)
2. 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)

Growth Rate(

−1 In order to generate pris\_gro, you will need to:

* + Sort your dataset by year [Hint: use sort year].
  + Use pris[\_n-1] for *xt*−1 in the formula above.

What is the average of the new variable pris\_gro?

1. 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)
2. 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)
3. 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.