

Homework 1

Eco 231 - Undergraduate Econometrics

Prof. Carolina Caetano

1. It is reported that average U.S. person nowadays has spent more on medical and health-related products/services, compared with the previous decades. Suppose a researcher is interested in examining which demographic or socio-economic group is identified to have higher medical/health expenditures than others. The researcher asks you with finding a suitable data set for this topic. One suitable data set may be the Consumer Expenditure Survey (CE). Try to find the data through the data sources talked about in the Stata lecture. You will need to locate the 2015 Public-Use Interview Survey Data Dictionary (in pdf file) to answer the following questions.
 - (a) Briefly describe the CE survey. Make sure you mention the survey's objectives, the target population and the methods used for data collection (you can also look at CE data webpage to answer this part).
 - (b) Provide variable name (at least two) that can be used to identify a person's demographic status (e.g, age, gender, race, education, etc.). Do the same analysis for the socio-economic status (e.g, income, assets).
 - (c) How would you identify total medical and health expenditures in the survey? Does the survey contain specific amount spent on each of these categories?
 - (d) Does the survey also collect any information related to health insurance? Is there any question asking whether the premiums are paid through payroll deductions?

2. You are interested in researching the link between regular exercise and mental health disorders (e.g., anxiety, depression, post-traumatic stress, etc.). Someone recommended you use the Behavioral Risk Factor Surveillance System (BRFSS) survey conducted by Center for Disease Control, which is available on the CDC website. You will need to locate the 2015 BRFSS Annual Survey Codebook (in pdf file) to answer the following questions.
 - (a) Briefly describe the BFRSS survey. Make sure you mention the survey's objectives, the target population and the methods used for data collection (you can also look at BRFSS main webpage to answer this part).
 - (b) What is the sample size of the survey? What proportion of survey interviewees completed the questionnaire in its entirety?

- (c) What is the proportion of males in the sample? What fraction of the respondents has completed 4-year college or more?
 - (d) Is there a question that asks respondents' general health status? Provide SAS variable name of that question. Report the proportion of respondents with invalid responses to that question (invalid includes: Don't know/Not Sure/Refused/Missing).
 - (e) List and describe the variables you may require to perform your estimation.
3. Now imagine that a researcher is interested in studying how an increase in police officers assigned to schools affects the level of school crime and a way that school responses to the crime. She is not sure if relevant data would be available on such an issue and hires you to help out. Briefly answer the following questions. We may penalize answers that are unnecessarily long.
- (a) Write down one scientific question related to the above research topic.
 - (b) Can you find a data set that might be useful in answering your questions? Give the name, source of the data (ICPSR might be a nice place to start looking), and the link.
 - (c) Briefly describe the data you found in part (b) and discuss key variables which can potentially help answer the questions you outlined in part (a) (No more than 10 lines).
4. This problem asks you to use the STATA file, `household.dta`, which is available for download on the website. It is public-use data from the household survey, and includes basic information of the household that participated in the survey. Our researcher is interested in the effects of wife's education attainment and labor market experience on wife's earnings in the labor market. Answer the following questions using STATA.
- (a) First, take a tour of the entire sample by using `browse` command. Are there any string variables in the dataset?
 - (b) What is the mean age of the wife in this sample? What are the 25th and 90th percentile of the distribution of age? What do these numbers mean?
 - (c) Draw a histogram of the wife's age distribution in the sample. When drawing a graph, make sure you have **15 bins** in this histogram. Describe the plot.
 - (d) What is the mean years of education for the wife in this sample? What is the median (50th percentile) of the distribution?
 - (e) What fraction of wives in the sample has 12 years of education attainment? Are 12 years the most common education attainment among wives?
 - (f) What is the mean years of wife's previous labor market experience in this sample? Draw a histogram over the wife's labor market experience to see how the distribution looks like. Does the distribution look symmetric? Describe the plot.

- (g) What percentage of wives has never had any past labor market experiences (0 years) at all?
- (h) The researcher wants to create a new variable, named `wedua1`, that has value 1 if wife's years of education attainment *exceeds* 12 years, and 0 otherwise. Use the method(s) introduced in the Stata lecture, create `wedua1` as described here.
- (i) Now, `wedua1` has value 1 if the wife completed at least 1 year in the college. To make this more explicit, create your own label of the variable `wedua1` that indicates such feature (any explanation that states the characteristics of `wedua1` is fine).
- (j) The researcher is able to compare mean earnings of wives who completed at least 1 year in college vs. mean earnings of wives who didn't attend any college. Use the method(s) introduced in the Stata lecture and `wedua1`, produce a table which allows you to compare these two. Make sure to use the variable `wage` (wife's 1975 average hourly earnings) for your calculations.
- (k) The difference of mean earnings between two groups is often referred to as "College Premium" measured in realized earnings. Do you see any college premium among wives in our sample? How large is that? Explain.