Problem Set 1

Statistics 506, Fall 2017

Due: Tuesday October 3 via Canvas

Course Homepage (https://jbhender.github.io/Stats506/)

Instructions

- You should submit your Stata code as text format files.
    - For problem 1, submit one file per part.
    - For problems 2 and 3 submit one file per question with the parts clearly labeled using comments.
    - Also submit a single PDF file containing your answers to all the questions that are posed below.
- Your code should be clearly written and it should be possible to assess it by reading it. Use appropriate variable names and comments. Your style will be graded using the style rubric (./StyleRubric.html).
- All work for this probem set should be done using Stata. Do not do any data input/output or pre/post processing of results using tools other than Stata.
- Some of these exercises will require you to use Stata commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Part of the purpose of these exercises is to be resourceful and self sufficient. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.

Problems

1. For the first problem you will work through online Stata turorials. As you work through the tutorials, write a script documenting the techniques and commands learned. Each script should clearly state and link to the tutorial being documented. sIn the file containing your answers, provide links to the tutorials you completed and the names of the scripts documenting the examples.
    a. **Converting between long and wide data formats.** First, read this blog post (http://www.theanalysisfactor.com/wide-and-long-data/) about wide vs long format data. You may also want to read the Stata help for the `reshape` command. Next, visit this page https://stats.idre.ucla.edu/stata/modules/ (https://stats.idre.ucla.edu/stata/modules/) and find the links 'Reshaping Data from Wide to Long' and 'Reshaping Data from Long to Wide'. Work through the examples and write a script (`.do`) documenting your work. Use comments to clearly organize the script into parts and provide brief explanations about what each reshape call is accomplishing.
    b. Visit https://stats.idre.ucla.edu/other/dae/ (https://stats.idre.ucla.edu/other/dae/) and choose and any **one** Stata example to work through and document.

c. Repeat part (b) for a second tutorial. Optionally, you may choose a tutorial from another source.

Here are copies of the data sets used in the examples:

- kids.dta (./kids.dta) or https://stats.idre.ucla.edu/stat/stata/modules/kids (https://stats.idre.ucla.edu/stat/stata/modules/kids)
- dadmomlong.dta (./dadmomlong.dta)
- dadmomwide.dta (./dadmomwide.dta) or https://stats.idre.ucla.edu/stat/stata/modules/dadmomw (https://stats.idre.ucla.edu/stat/stata/modules/dadmomw)

My thanks to the student who provided the links.

2. Use the RECS data discussed in class to answer this question. Produce nicely formatted tables or graph to justify your answers.

   a. Which state has the highest proportion of wood shingle roofs? Which state(s) the lowest?

   b. Compute the proportion of each roof type for all houses constructed in each decade. Which roof type saw the largest relative rise in use between 1950 and 2000?

3. For this question you will analyze data from the 2009 NHANES survey. You may ignore the survey weighting in your analysis and treat this as a simple random sample.

   a. Download the Audiometry data (AUX_D) from here (https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&CycleBeginYear=2005) and the demographics file (DEMO_D) from here (https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Demographics&CycleBeginYear=2005). Determine how to load them into Stata and then merge on the common identifier `seqn`. Drop all cases without audiometry data.

   b. Read the information under the headings 'Eligible Sample' and 'Protocol and Procedure' in the doc file for the audiometry data at the first link above. For each hearing threshold test, compare the old and young sub-populations stratified by left and right ear. You may compare each frequency separately and only need to use the first test at each frequency. Based on these data, is hearing loss due to age more common at particular frequencies? Is either left or right ear more prone to hearing loss? Produce a nicely formatted table or graph to justify your answer.

   c. For this question you should consider only the right (not left) ear and can again analyze each frequency separately. Is either gender more prone to age-related hearing loss?

Solutions

- Problem2_final.do (./Problem2_final.do)
- Problem3_final.do (./Problem3_final.do)