SOC 5050: Problem Set 03

Christopher Prener, Ph.D.

September 19th, 2016

Directions

Please complete all steps below. Include your research log, final stack of do-files, log-file, plots, and markdown file. Your markdown file should contain source code for PART 3 as well as your narrative, which should contain answers to the questions in PART 3. You should also include a scan of your solution to PART 1. All requested documents should be uploaded to your GitHub assignment repository by 4:20pm on Monday, September 26th, 2016.

Part 1: Bayes' Theorem

1. Neighborhood "A" is has a reputation for violent crime, though it has been going down in recent years. The police department estimates that, given this trend as well as a trend where drug arrests have been increasing over the past few months, the probability of the violent crime rate continuing to fall is 20%.

   The City Council is interested in investing in cameras to monitor high crime areas. You are asked to evaluate the potential impact of these cameras on the violent crime rate in neighborhood "A". You hypothesize that the introduction of cameras leads to a higher probability of a decrease in the violent crime rate (i.e. the condition).

   After conducting a literature review, you find two estimates of the impact of cameras in other cities. The first is that the probability of violent crime rates decreasing given the installation of cameras is 65%. However, other studies caution that in 12% of neighborhoods, the cameras have no impact on violent crime.

   (a) What is the prior probability (x) given in the scenario above?
   (b) What is the probability of the condition if the hypothesis is true (y)?
   (c) What is the probability of the condition if the hypothesis is false (z)?
   (d) Estimate the posterior probability that the violent crime rate will continue to fall given the introduction of cameras in neighborhood "A". Use the formulate given below and include your

work done by hand. Also use the `display` command in Stata to check your work.

$$\frac{xy}{xy+z(1-x)} \tag{1}$$

Part 2: Initial Steps

2. Open Atom and create a new Markdown file (change the document's language to `GitHub Markdown`). Use this as your research log for this Problem Set. Take careful notes about your plan, organizational steps, and execution steps for each aspect of this problem set.
3. Download the dataset for this Problem Set from this link.
4. Un-zip the dataset and move the files into your folder hierarchy for the course.
5. Open the file `33041-descriptioncitation.pdf` and read about the dataset you have downloaded.
6. Next, open the codebook (`33041-0001-Codebook.pdf`) and skim through it. Look for information about what types of information is included in the dataset and how missing data are currently being handled.

Part 3: Data Cleaning

7. Beginning with the appropriate course template, complete a master do-file to trigger your `data.do` and `analysis.do` files.
8. Beginning with the appropriate course template, construct a data do-file that completes the following tasks:
   (a) Remove variables from your dataset[1] so that you only have the following variables remaining: `CASEID, METRO, AGECAT, SEX, RACE, YEAR, QUARTER, DAYPART, DRUGID_1, DRUGID_2, DRUGID_3, DRUGID_4`, and `DRUGID_5`.
   (b) For each of the variables you retained in the previous step, properly declare values as missing. If applicable, create missing values that differentiate between different reasons data may be missing. Be sure to use the full workflow for recoding variables.

[1] *Hint*: See Week 03's lecture materials.

(c) Starting with the variable `METRO`, create a new, recoded variable that is a *categorical* measure representing four major regions of the country: east, southeast, midwest, and west. Declare respondents residing in "other" metro areas as missing in this new variable. Be sure to use the full workflow for recoding variables.

(d) Starting with the variable `AGECAT`, create a new, recoded variable that is an *ordinal* measure presenting a smaller set of age categories. Select categories that you believe make the most sense with the goal of having four or five values in your new variable. Be sure to use the full workflow for recoding variables.

(e) Starting with the variable `RACE`, create a new, recoded variable that is an *binary* measure representing non-white individuals. Be sure to use the full workflow for recoding variables.

(f) Starting with each of the `DRUGID` variables, create new, categorical variables that represent a number of specific substances. The categories that should be included are: `ALCOHOL (ETHANOL)`, `COCAINE`, `HEROIN`, `MARIJUANA`, `METHADONE`, and `METHAMPHETAMINE`.[2] All other substances should be properly coded as missing.

[2] *Hint*: You will need to use the Data Browser to identify the numeric values for each of these categories.

Part 4: Missing Data and Descriptive Statistics

9. Beginning with the appropriate course template, construct a analysis do-file that completes the following tasks:

   (a) Conduct a missing data analysis on all of the variables you modified or created in PART 3. How much missing data is there? Are these data possibly MCAR or MAR? Are there threats to generalizability that are revealed by these tables?

   (b) For each variable you created or modified in PART 3, create a frequency table and obtain the appropriate descriptive statistics. Describe your findings and include a justification for why these are the appropriate statistics to report.

   (c) For each variable you created or modified in PART 3, create an appropriate plot (or plots), interpret the plot, and include a justification for why this is the appropriate plot.

Grading Rubric

Since this is a combined lab and problem set, the entire assignment is worth 36.25 points.

Part 1 Completing question one correctly is worth three points.

Part 3 This section is worth ten points towards the 36.25 point total for the assignment. These points are awarded based on correctly completing all the workflow steps for recoding missing and valid data.

Part 4 This section is worth ten points. A third of the credit comes from the proper use and execution of the relevant Stata commands, a third comes from stating the correct answers, and a third comes from your accompanying justification.

Stata Do-File The overall quality of the Stata do-file stack is worth ten and a quarter points. This grade will be based on the clarity, organization, and layout of your do-files.

Design An additional three points are based on the layout and design of each of your figures. This grade will be based on the use of schemes as well as customization of the plots (titles, subtitles, and notes).

Document Details

Document produced by Christopher Prener, Ph.D. for the Saint Louis University course SOC 5050 - QUANTITATIVE ANALYSIS: APPLIED INFERENTIAL STATISTICS. See the course wiki and the repository `README.md` file for additional details.