Assignment 2: Data

January 23, 2020

The goal of this exercise is to familiarize you with large and realistic data, and learn basic techniques for data manipulation and descriptive statistics. The data is coming from a field experiment in Ghana, which provided information to students about schools. The main dataset is available here.

The data comes from a project that aims to identify the impact of information on demand for education. The project provides a randomly selected group of Ghanaian junior high school students with application strategies and information about the selectivity and exam performance of secondary schools. Hence, changes in subsequent secondary school application and enrollment can be studied. Students are asked to make several choices, which are usually suffixed with a number. For example, variables with "choices_1" in the name refer to characteristics of the first choices.

Exercise 1 Basic characteristics

- Find the number of rows and columns. Using the function summary, sink descriptive characteristics of all variables into a file called "summary.txt".
- Find an individual identifier, which is defined as the variable that uniquely identifies each row.

Exercise 2 Small bases

The goal of this exercise is to create smaller datasets. Hint "grep".

- Create a dataset to be called dat.surv which includes all variables that start with the keyword "stud_base" plus an individual identifier.
- Create a dataset to be called dat.admin which includes all variables that start with the keyword "alladmin" plus an individual identifier.
- Create a dataset to be called dat.rest which includes all variables that did not start with "alladmin" and "stud_base" plus an individual identifier.

Exercise 3 Consistency

- Find gender variables in each of the small data(s). Are they consistent?
- Find age variables in each of the small data(s). Are they consistent? Count the number of students with inconsistent age variables.
- The survey asks three questions:
  - What is the best aggregate BECE score you think you might get, if the BECE exams go very well? (range 6 to 54)
  - What is the worst aggregate BECE score you think you might get if the exams do not go well?
  - What aggregate BECE score do you think you are most likely to get on the BECE this year? (range 6 to 54)

  These questions can be found by searching keywords "BECE" and words like "best", "worst", and "likely".

Document these following events.

- Inconsistency: Worst outcome better than the best outcome.
- Over-confidence: Likely outcome better than the best outcome.
- Under-confidence: Likely outcome worse than the worse outcome.

Exercise 4 Balance

- The data contains a treatment variable that may recovered searching for the keyword "treat".
- Check whether variables such as gender, age and SHS region are balanced across treatments.

Exercise 5 Recoding and histogram

- The survey includes a question, "What is the highest education level you want to complete?". The levels of this variable are
  - Junior high school
  - Technical or vocational training
  - Senior high school
  - Nursing or teacher training
  - Polytechnic
  - University.

The variable can be found by search keywords “educ” and “want”. Recode this variable to reflect the label provided in the question.

- Plot the histogram of desired education by gender.

Exercise 6 Manipulating the data

The data elicits individual preferences over schools. For example, there is a question: “ Now imagine if you were making the decision yourself, and you could select any choices based on how much you like them, without worrying about whether you could gain admission to the schools, which schools and programs would you list. Those variables can be recovered using the keyword ”mychoice”.

- Create a dataset that consists of variables mychoice and associated with academic tracks using“pgm”. Compute a frequency table of programs by ranked choice number.
- Create a dataset that consists of variables mychoice and associated with regions using“region”. Compute a frequency table of regions by ranked choice number.
- Individuals are also asked What is aggregate score do you think you would need for admission to this choice? This variable can be recovered using the keywords mychoice and “bece”. Plot the average, sd, the 25% quantile and 75% by ranked choice. Randomly select 5 individuals, and report how the expected score changes by ranked choices. Create a dummy variable “reverting” which is equal to 1 if the aggregate score are decreasing by ranked order list or 0 otherwise.

Exercise 7 Probit

Run a probit of the dummy variable reverting on measure of consistency, overconfidence, underconfidence as well as gender, age, and expected education. $Hint : glm$.