Busara Center for Behavioral Economics
Data Analyst – Data Task
Due: Saturday Midnight, 27th April 2019

Instructions

All necessary files for completing this test are included in the attached folder. Ideally you will use R, if so send a knitted HTML or PDF document with the code, outputs and notes explaining the output. If you do not have access to R please use a software of your choice to complete the task, including all supporting files, final datasets and any necessary written instructions to replicate your code.

Please note, we do not recommend you use SPSS as we will have difficulty assessing your ability to code (we prefer the use of Stata or Python if you do not have access to R).
You may consult any resources you like, except for other people. Please list any resources that you consulted in the analysis report which could be an R Markdown if you used R. If at any point you are stuck, explain (preferably commented in your output file) what you would have done had you had more time or knew the correct commands for doing it.
Try to get through as much of the test as you can in the time allotted; even if your answer depends on previous steps that you were unable to do, you will still get points for demonstrating that you would have gotten the correct answer if you had successfully completed all previous steps.

Question 1

1.1 Description of the dataset:

The government of Uganda rolled out a school feeding program in selected districts to provide primary school pupils with a free meal on school days with the hope of increasing attendance rates in public primary schools. Schools in each district were randomly assigned to treatment and control groups. Your task is to assess the impact of the programme on school attendance. The Excel workbook contains two worksheets; one with the data on attendance and the other contains the district names. Each row of data contains data for each school within the district.

1.2 Problems

a) Import the "**School data.xlsx**" file and compute the total number of student enrolled in each school.
b) Create the school ID variable by first sorting the data within each district by the total number of enrollees per school. Let the ID be 1 for the school within each district with the highest number of enrolled students, 2 for the second highest and so on.
c) Add district names to the main dataset using data from the second sheet. Make sure all towns are named and drop any irrelevant towns.
d) Check the numeric variables for outliers. What is the importance of this exercise? Explain how you handled outliers if there were any? If they weren't give a brief explanation of how you would have handled them if they were present.
e) Label values for the treatment variable appropriately (1 = Treatment, 0 = Control).
f) Create a well labelled graph that shows the difference in female attendance between treatment and control schools.

g) Create a function that takes in a numeric variable and outputs a table of summary statistics and a histogram showing the distribution of the variable. Test out the function using variables in this dataset.
h) Regress total attendance on treatment, with district fixed-effects.
i) Regress total attendance on treatment, with district fixed-effects and controlling for the total number of enrolled students at each school.
j) Explain the statistical methodology used and reasons for the choice. Give a detailed description of the analysis results and why there are differences, if any, between model in (h) and (i). When displaying the table of outputs, leave out the district fixed effects.

Question 2

2.1 Description of the dataset:

The "**HealthInsurance.xlsx**" is a cross sectional dataset originating from a Medical Expenditure survey conducted in 2010 in some African country. The main objective of the study was to understand the health insurance uptake and also find out factors that influence this uptake.

2.2 Problems

a) Import the dataset and clean the data of any anomalies.
b) What is the distribution of the sample in terms of (i) gender and (ii) age?
c) Ideally, we expect respondents with poor health status to have health insurance. Is this the case according to this data? Using an appropriate statistical test, investigate the relationship between the two variables. Specify the hypothesis being tested here.
d) Determine the ethnicity of majority of the respondents. Filter respondents for the major ethnicity and count of respondents who are self-employed and have a health insurance cover.
e) An insurance company would like to use this data to understand what factors drive insurance uptake. Select an appropriate statistical model for this task and give reasons for the choice.
f) Fit the model on the data and construct 95-percent confidence intervals for the coefficients of the explanatory variables. What conclusions can you draw from these results? Present regression output in the most professional way and add any notes you think are necessary to explain the model specifications and output.
g) Use a likelihood-ratio test to test the null hypothesis that none of the explanatory variables influences insurance uptake.

Busara Center for Behavioral Economics
Data Analyst – Data Task
Due: Saturday Midnight, 27th April 2019

Deliverables

Name the folder containing your output using this protocol (Surname_Firstname_Busara_April_2019). Please zip and email back the following:

- All the raw data files
- A Stata do file and word/PDF document explaining the analysis process and output [If you used Stata]
- An. Rmd file and a knitted R Markdown document HTML/PDF format [If you used R]
- A Jupyter Notebook [If you used Python]

Note: The scripts files should be replicable: Another person should be able to run it and produce the exact same results on their computer. All the best.