Final Project
STAC53F: Applied Data Collection
Due on 19 December, 2020 11:59 PM EDT Sharp in Quercus

Final Project: The final project is due on **19 December, 2020 11:59 PM EDT** and consists of a data analysis on a dataset. The deadline will be strictly applied. At no circumstances students can submit late. Please make sure that you start the submission process early so that your project is graded.

Students will be required to demonstrate their understanding of the methods based on course materials by developing a reasonable sampling techniques and regression model using the techniques taught in class. The students will be responsible for choosing appropriate methods to apply and providing appropriate justifications defending their choices.

The final project will be done individually, and must be typed and submitted by the stated deadline. The project needs to fulfill the following criteria:

- Font: 12-point font in a style similar to Times New Roman
- Spacing: single-spaced
- The word limit for the final project is **1200**. This excludes the title page, table/figure captions and appendix.
- Maximum 5 tables/figure will be allowed in the project report. The tables and figures should be relevent, should convey the purpose of the project. All tables and figures should have captions. you may use any combination of tables and figures
- Up to 3 additional tables/figures but they should only be included if they are relevant to the analysis and are referred to in the main text.
- You must submit the report in a standard file format (e.g., .doc, .docx or a pdf).
- Please provide all your R codes in Appendix of the report.

In order to pass the course, you must submit the final project.

For this problem you need to load the NHANESraw dataset using the following command

```
library(NHANES)
library(tidyverse)

## Create the data ##
small.nhanes <- na.omit(NHANESraw[NHANESraw$SurveyYr=="2011_12"
& NHANESraw$Age > 17,c(1,3,4,8:11,13,24,25,61,77)])
small.nhanes <- small.nhanes %>%
group_by(ID) %>% filter(row_number()==1)
```

This is data collected by the US National Center for Health Statistics (NCHS). To check the variable description please type `?NHANES` in `R`. The preceeding codes create a small subset of the original NHANESraw dataset. The original dataset has 76 variables. The `small.nhanes` dataset has 17 variables. We have only selected data from people with age > 17 years.

With this dataset answer the following questions, randomly select $n = 500$ observations from the data (i.e., perform a simple random sampling). For this selection use your student ID as the seed. Then perform the following tasks,

1. `BPSysAve` is the outcome of interest and `Smoke100` is the exposure. Identify the relationship between the variables. Perform the univariate and bivariate analyses. That is produce some plots to visualize the variables and their relationships and then perform appropriate statistical tests.

2. Perform an appropriate regression model for the outcome `BPSysAve`. For this model please consider all variables except the `ID` variable. However other than `Smoke100` you can only choose maximum of 3 variables. Please explain the reasons which variables you have chosen.

3. For the next step use stratified random sampling. Here `SDMVSTRA` is the stratum variable. Here `SDMVSTRA` are the strata and there are 14 strata here. The cost of sampling from each stratum here is as follows, $C = 52, 50, 46, 53, 48, 48, 47, 57, 53, 47, 54, 40, 43, 44$. Perform the following tasks. Calculate the the stratum specific sample size $n_j; j = 1, 2, ..., 14$ and the overall sample size n. For the calculations consider the following,

   - Here N_j is the total number of observations for stratum j
   - The Margin of error $ME = 4$ and the level of significance $\alpha = 0.01$.
   - The stratum specific variance σ_j can be calculated by dividing the stratum specific range of the blood pressure (`BPSysAve`) by 4. That is calculate $\sigma_j = Range_j/4$.

4. Now based on the n_js you calculated from (a), collect a sample (without replacement) from the `small.nhanes` dataset from each stratum. Then perform an appropriate regression model for this design. Consider the same set of variables those you used for (2).

Compare the results you obtained from (2) and (4). Explain your findings. The final project will be submitted as a project report, which consists of:

- **Introduction section**: where you introduce the purpose and relevance of the project. You can also include some literature review on the `NHANES` dataset if applicable.

- **Methods section**: Please describe and explain the methods, tools and techniques used to arrive at your final model here. Need to show some exploratory data analysis.

- **Results section**: here you present a description of your study sample, important results that led you to make crucial decision in building your model, and the final model and any other important results

- **Discussion section**: here you interpret your final model and describe why it answers the research question and why it is important, as well as discuss any limitations that still exist based on your results.

ALL THE BEST!