

Final Project
STA302H1F/STA1001HF
Due on 14th June, 2021 11:59 PM Sharp in Quercus
All relevant work must be shown for credit.

Final Project: The final project is due on **June 14, 2021 by 11:59PM EST** and consists of a data analysis on a novel 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 regression model using the techniques taught in class. The students will be responsible for choosing the correct 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:

- To submit your results, you will be required to prepare a 5 minute presentation that you will need to record (using your computer, phone, etc.). You will be required to display your T-card alongside your face at the beginning of your video to verify your identity.
- You will be required to submit the R codes that you have created in a separate file. This helps us to check the reproducibility of your codes.
- You will need to display the results of your project in a logical way using slides (e.g. Power-Point, latex, R Markdown or other) and record yourself discussing these results, with a focus on why you chose to do certain things and interpretation of your results for non-statisticians.
- A rubric will be provided shortly within this week.
- Presentations should be submitted on time (i.e. by the deadline).
- **There are no make-up final projects.** A missed final project will be given a grade of 0.

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

```
## If the package is not already installed then use ##
install.packages('NHANES') ; install.packages('tidyverse')
library(tidyverse)
library(NHANES)
small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2011_12"
& NHANES$Age > 17,c(1,3,4,8:11,13,17,20,21,25,46,50,51,52,61)])
small.nhanes <- as.data.frame(small.nhanes %>%
group_by(ID) %>% filter(row_number()==1) )
nrow(small.nhanes)
## Checking whether there are any ID that was repeated. If not ##
## then length(unique(small.nhanes$ID)) and nrow(small.nhanes) are same ##
length(unique(small.nhanes$ID))
```

This is data collected by the US National Center for Health Statistics (NCHS). To check the variable description please type `?NHANES` in R. The preceding codes create a small subset of the original NHANES 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 500 observations from the data. For this selection use your student ID as the seed (you can follow the next chunk of codes for this). This is the training set. The rest of the data will be used as a test set. The test set should not be used for model fitting and validating at any point during the analysis of the project.

```
## Create training and test set ##
set.seed(1002656486)
train <- small.nhanes[sample(seq_len(nrow(small.nhanes)), size = 500),]
nrow(train)
length(which(small.nhanes$ID %in% train$ID))
test <- small.nhanes[!small.nhanes$ID %in% train$ID,]
nrow(test)
```

The combined systolic blood pressure reading (`BPSysAve`) is our outcome of interest. Every other variable other than the ID can be considered as predictors. We are mainly interested on the effect of smoking (`SmokeNow`) on the combined systolic blood pressure reading. However, we are also interested in the prediction of the combined systolic blood pressure reading and identifying which variables are the best for the prediction. Based on the data analysis techniques you learned from this course perform a complete analysis on the dataset. Your analysis should include (but is not limited to):

- Model Diagnostics
- Checking for the variance inflation factor (VIF)
- Variable selection
- Shrinkage methods
- Model Validation
- Checking the prediction error on the test set after applying various model selection techniques
- After selecting the best model interpret and explain the parameter estimates
- Conclude on the effect of predictors on the combined systolic blood pressure reading

However, you have to justify the aforementioned methods and have to use them accurately.

The final project will be submitted as a presentation. However, to structure your presentation please present in the following order:

- **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 and if you have some time.

- **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 (graphs and tables).
- **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!