STAT 6333: Statistical Learning
Case Study 1 Instructions
Fall- 2021
Due: October 16th, 2021

Expectations of the Case Study Analysis:

The main goal of this case study is to get hands on experience in applying the simple/ multiple linear regression and k-means regression methods to a real-world data set.

Dataset Details:
Tinnitus is one of the most distressing hearing-related symptoms. This can be a significant problem that negatively impacts the quality of life. An internet cognitive behavioral therapy (referred to as the treatment here in the data set) intervention has been developed in UK to improve access to evidence-based tinnitus treatment. The Tinnitus Functional Index (TFI_score) has been used as the primary assessment measure to quantify tinnitus distress prior and after the treatment.

The dataset ("**CaseStudy1.csv**)" contains the pre and post TFI_scores, clinical and demographic factors related to a pre-post interventional study of 142 subjects with Tinnitus. The specific column details are as follows.

Column	Name	Description
A	Subject_ID	Subject ID of the participant
B	Group	Subject's Treatment/ Control group details
C	HHI_Score	Hearing survey- Overall score- 0-40 (higher score more severe)
D	Generalized Anxiety Disorder (GAD)	Anxiety sum: 0-21 (higher score more severe)
E	Patient Health Questionnaire (PHQ)	Depression sum: 0-28 (higher score more severe)
F	Insomnia Severity Index (ISI)	Insomnia total: 0-28 (higher score more severe)
G	Satisfaction with Life Scales (SWLS)	Overall score, satisfaction with life, like Quality of Life (QOF). Higher scores better QOL (opposite to all other scales)
H	Hyperacusis	0-42 (higher score more severe)
I	Cognitive Failures (CFQ)	0-100 (higher score more severe)
J	Gender	1-Male, 2- Female
K	Age	In years
L	Duration of tinnitus	In years
M	Pre TFI Score	TFI score at the beginning of the study: Tinnitus score out of 100, higher more severe
N	Post TFI Score	TFI score after the completion of the study; Tinnitus scores out of 100, higher more sever.

While performing the data analysis, for each analysis method, you are expected to comment on the following.

Multiple Regression Analysis

1. Explore the data set by performing any descriptive analysis. This includes creating pie charts, histograms, correlation analysis, etc. as necessary.

2. Please check whether there are any missing values in the Pre and Post TFI scores. If there are any such missing values, use the data imputation with mean (corresponding to that Pre or Post TFI score) to perform the data imputation. After the data imputation please create a new variable called "**TFI_Reduction**" by subtracting "**Post_TFI_Score**" by "**Pre_TFI_Score**". TFI_Reduction will be used as the response in our multiple linear regression model.

3. If you see any other numerical measurement with missing values, please use mean to impute the data. If there are any missing values for any categorical measurement, please use "mode" to impute the data.

4. Partition that data set (obtained at step 2) into a train (80% of the data) and a test set. (Hint. Use *set.seed(123)* and *sample()* in base package or *createDataPartition()* in Caret package)

5. Perform the multiple regression analysis on the data. Use the "TFI_Reduction" of each subject as the response. Comment on your findings. (You may use best subset selection/forward/backward selection methods to select the best multiple linear regression model with *lm()* in R).

6. Once you select a best model with high prediction power (use multiple metrics like Adjusted R2, AIC and BIC to select the best model), perform the model diagnostics. If you see any violations in the model assumptions, please take the appropriate actions to correct them. (For example, if you see a U shape pattern in the residual plot, try including a quadratic term, If you see any potential influential point, create two regression models, both with and without that data point to evaluate how the regression estimates and their standard errors get impacted). Comment on what you saw and on the actions that you took to justify your approaches.

7. After you clarify that there is not any issue with the model assumptions, use that model to find out the factors which highly influence the reduction in TFI score. Comment on your findings.

8. Make the predictions on the test data set. Comment on the mean square error on the testing data set.

9. Perform ridge regression, lasso regression, principal component regression, and partial least squares regression to do the same previous steps to decide which multiple linear regression has the highest predictive power via the smallest testing error.

K-means Regression

9. Use K-means regression to train several regression models after selecting the best k value, using Monte-Carlo cross validation.

10. Make the predictions on the Testing data set and obtain the mean square error.

11. Compare your best test mean square error with mean square error that you obtained in the multiple linear regression (with all of the different methods). Which method gives the lowest mean square error, Multiple linear regression or K-means regression?

After the completion of the analysis, you need to hand over a small report (with a minimum of 5 pages) with your findings. Make sure to attach your R code at the end of the report.