Middle East Technical University
Department of Statistics
2020-21 Fall Semester
STAT 467 Multivariate Analysis

HOMEWORK 2

(Due: December 22, 2021, Friday – 23.59)

- *You should work on your own. Please feel free to get help from me, but not from anyone else. Let me know if my wording in the questions is not clear. Therefore, absolutely, no late homework will be accepted.*
- ***I will open a forum on ODTUClass. Please write your questions in this forum.***
- *Please use R Markdown to do your homework. Then, produce a* ***word (then convert it into pdf)*** *or* ***html*** *file using it. If you have any question or problem, please let me know.*
- *You will submit your homework as a zip file including your pdf or html file.*

Read insurance.csv data set.

The explanation of the columns in the dataset are given below.

- **age:** age of primary beneficiary
- **sex:** insurance contractor gender, female, male
- **bmi:** Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,objective index of body weight (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9
- **children:** Number of children covered by health insurance / Number of dependents
- **smoker:** Smoking
- **region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges:** Individual medical costs billed by health insurance

Please answer the following questions using R-Studio.

a) Do both BMI and Charges follow the normal distribution?

Create a subdata including only BMI and Charges, then apply the multivariate normality test. If you detect non-normality please list the alternative solutions. (Don't have to apply them!)

b) Do both BMI and Charges vary with respect to smoking? Please state the name of the method you use.

c) Apply K-Means Clustering to cluster your observations. Does the optimum cluster number equals to number of region in the data? Draw the clustering plot.

Hint: Before starting the procedure, please run the following code to prepare your dataset. (Assume that you call insurance.csv as data)

Ozdemir, O.
users.metu.edu.tr/ozancan
e-mail: ozancan@metu.edu.tr

```
> data<-read.csv("insurance.csv")
> data$sex_n<-ifelse(data$sex=="female",1,0)
> data$smoker_n<-ifelse(data$smoker=="yes",1,0)
> data1<-data[,-c(2,5,6)]
```

Consider `USairpollution` data set in `MVA` package. Then, please answer the following question.

d) Divide your dataset into two part. Please use **data1,** created in the previous part, for the process. (80% Train, 20% Test)

e) Fit a linear regression where you consider BMI as your response variable. Please be sure that all variables in your model must be significant. Then, make a prediction by using your test data and calculate RMSE for your model.

Hint 1: Please use **data1 train** and **data1 test** for the regression process.

Hint 2: You can use backward elimination to reach the model having significant predictors.

Hint 3: Use `predict` function to make a prediction from the linear model.

Hint 4: You can use this code to calculate RMSE

```
> rmse<-sqrt(mean(y_hat-test$y))
```

f) Fit a decision tree with optimal parameters and calculate RMSE for it.

Hint 1: Please use **data1 train** and **data1 test** for the process.

Hint 2: You can use the codes in Recitation 10 and/or Exercise 1 in Recitation 10.

g) Construct a bagged tree by using `caret` package and show OOB RMSE.

Hint 1: Please use **data1 train** for the process.

Hint 2: Create a bagged tree with 10-fold cross validation.

h) Please state the model name having the smallest RMSE value.