STAT 40001/STAT 59800 **Statistical Computing** **Fall 2020**

Homework 5

Name:

Due : November 24, 2020 PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output containing ERRORs). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where x is the year coded and y is the sales in thousands of units:

x_i:	0	1	2	3	4	5	6	7	8	9
y_i	98	135	162	178	221	232	283	300	374	395

a) Prepare a scatter plot of the data.
b) Fit a simple linear regression model and perform the residual analysis.
c) Use Box-Cox procedure to find an appropriate transformation of y.
d) Fit a simple linear regression for for the transformed data.
e) Compare the models before and after the transformation.

Q.N. 2) Descriptive information and the appraised total price (in Euros) for apartments in Vitoria, Spain are available in the data set `vit2005` in `PASWR` package.
a) Access the data set and state its dimension.
b) Print the list of the variables included in the data set.
c) Develop a parsimonious multiple linear regression model that can be used to predict the total price (`totalprice`) of apartments by their hedonic (structural) characteristics `area, age, floor, rooms, toilets, garage, elevator` and `tras`.

Q.N. 3) A study was conducted attempting to relate home ownership to family income. Twenty households were selected, and family income(x) was estimated along with information concerning home ownership (y = 1 indicates yes and y = 0 indicates no). The data are provided in Brightspace with this assignment.
a) Fit a simple logistic regression model for the subject data and display with the scatterplot.
b) What is the estimated probability that a family with an income of $45, 000 owns a house?

Q.N. 4) We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance. The response default falls into one of two categories, Yes or No. Access the data set `Default` from the `ISLR` package and fit a simple logistic regression model to model the probability of defaulting the credit card payment based on the credit card balance.

Q.N. 5) According to the web site `http://www.keepkidshealthy.com`, risk factors associated with premature births include smoking and maternal malnutrition. A birth is consider premature if the gestation period is less than 37 full weeks. Also note that the body mass index(BMI) can be used as a measure of malnutrition. Do you find this to be true with the data in babies provided in the `UsingR` package?
Tasks to perform:
a) Extract the variables of interest: gestation, smoking status, mother's height and weight, and birth weight of the babies.
b) Clean the data set as there are some missing values coded as 9, 99, or 999.
c) Calculate the BMI of mothers.
d) Create indicator variable(1 for premature and 0 for not premature) babies.
e) Fit a logistic regression model with `smoke` and `BMI` as a predictor variable and `premature` as a response variable.