MSiA-400 Everything Starts with Data
Lab Exercise #1

Due Date: Monday, October 16, 09:00 am

EXERCISE INSTRUCTIONS: Q1: submit R script and an example using data "Tensile.txt"; Q2 & Q3: submit short answers, related code and print for each problem if necessary. Push your answers to Github (required) and Canvas (optional).

Problem 1

In the lab session, we used aov() for one-way ANOVA analysis. Please write your own one-way ANOVA function in R which takes two vectors (i.e. data and group labels) as input and the F-test result (e.g. reject null hypothesis or not) as output. You don't have to output the entire ANOVA table.

Problem 2

Data set *bostonhousing.txt*, created by Harrison and Rubinfeld [1978], concerns housing values in suburbs of Boston. The attributes include

MEDV	Median value of owner-occupied homes in $1000's
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population,

in which MEDV is the response variable. The summary of the data set is below.

Name of the data set	bostonhousing
Number of observations	506
Number of attributes	14 (1 response variable and 13 explanatory variables)

Problem 2(a)

Build regression model reg and display summary() of the model. Pick two explanatory variables that are least likely to be in the best model, and support your suggestion in one sentence.

Problem 2(b)

Build regression model reg.picked by excluding the two explanatory variables selected in problem 2(a). Display summary() of the model.

Problem 2(c)

For a regression model, the mean squared error (MSE) is defined as $\frac{SSE}{n-1-p}$, in which p is the number of explanatory variables used in the model. The mean absolute error (MAE) is similarly defined: $\frac{SAE}{n-1-p}$. Display MSE and MAE for regression models reg and reg.picked from the previous problems. Based on MSE and MAE, pick one model you prefer.

Problem 2(d)

Run step() using regression model reg in problem 2(a). Compare the model with reg.picked in problem 2(b).

Problem 3

Import *labdata.txt*. The summary of the data set is below.

Name of the data set	labdata
Number of observations	400
Number of attributes	9 (1 response variable and 8 explanatory variables)

Column y is the response variable and remaining attributes x1,x2,... are the explanatory variables.

Problem 3(a)

Build regression model reg and display summary() of the model

Problem 3(b)

For each explanatory variable, plot it against the response variable. Based on the scartter plots, pick one variable that is most likely to be used in a piecewise regression model. Attach one plot associated with the variable you pick.

Problem 3(c)

Calculate the mean of the variable you pick in problem 3(b) and build piecewise regression model reg.piece using the mean. Is model reg.piece better than model reg in problem 3(a)? Support your argument in one sentence.

Reference

David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. Journal of environmental economics and management,5(1):81-102,1978.