**BU.510.650 Homework #1**

**Data Analytics Page 1 of 2**

**Spring 2021 The Johns Hopkins University**

**Arnab Bisi Carey Business School**

**Homework #1**

Due: 02/06/21, 11:59pm

1. (a) Run basic R codes in RStudio provided in the file “R-code-Topic-00.R” and submit your output. (1 point)

<R-code-Topic-00.docx>

(b) Download the two data files (Advertising.csv; pickup.csv) from Blackboard and run the R codes in R-Studio provided in the file “R-code-Topic-01\_introduction.R”, and submit your output here. (2 points)

<R-code-Topic-01_introduction.docx>

1. (a) Write the R-code to create a matrix in which the first column consists of x variable taking integer values from 1 to 6 and the second column taking 6 random values from a standard normal distribution. Provide the matrix you generated. (1 point)

mtrx <- **matrix**(**c**((1**:**6), **rnorm**(6)), ncol = 2, byrow = F)

mtrx

## [,1] [,2]  
## [1,] 1 0.3722903  
## [2,] 2 0.3296978  
## [3,] 3 -1.0543443  
## [4,] 4 1.7048610  
## [5,] 5 0.3097746  
## [6,] 6 -1.2673363

(b) What is the key difference between a matrix and a data frame in R? Provide an example to demonstrate the difference. (1 point)

Matrix: A two-dimension homogeneous data collection, of which the number of rows and columns are predefined and cannot be changed.

Data Frame: A two-dimension heterogeneous data collection, of which allows different data types to co-exist and has dynamic-allocated rows and columns.

**setClass**('Person', **representation**(name = 'character', age = 'numeric'))  
john <- **list**(**list**(**new**('Person', name = 'John Smith', age = 42)))  
  
mtx <- **matrix**(**c**(1**:**9), ncol = 3, byrow = T)  
mtx

## [,1] [,2] [,3]  
## [1,] 1 2 3  
## [2,] 4 5 6  
## [3,] 7 8 9

mtx[4,] <- **c**(10, 11, 12)

*# Error in `[<-`(`\*tmp\*`, 4, , value = c(10, 11, 12)) : subscript out of bounds*

mtx[1,1] <- john

*# Matrix structure compromised: variable mtx is now a list object*

**class**(mtx)

## [1] "list"

df <- **as.data.frame**(**matrix**(**c**(1**:**9), ncol = 3, byrow = T))  
df

## V1 V2 V3  
## 1 1 2 3  
## 2 4 5 6  
## 3 7 8 9

df[4,] <- **c**(10, 11, 12) *# Dataframe allows appending new row to existing collection*   
df[1,1] <- john *# Dataframe allows storing heterogeneous data*   
df

## V1 V2 V3  
## 1 <S4 class 'Person' [package ".GlobalEnv"] with 2 slots> 2 3  
## 2 4 5 6  
## 3 7 8 9  
## 4 10 11 12

3. Download the Auto data set, from the course Blackboard page to answer questions (a) – (e). Make sure that the missing values have been removed from the data.

(Total 5 points for parts (a)-(e))

1. Which of the predictors are quantitative, and which are qualitative?

Quantitative:

MPG, cylinders, displacement, horsepower, weight, acceleration, year

Qualitative:

Origin, name

1. What is the range and median of each quantitative predictor? (*Hint*: For range, use the range() function.)

quant <- **sapply**(auto[, 1**:**7], **function**(c) **c**(**range**(c),**median**(c)))   
**row.names**(quant) <- **c**('MIN', 'MAX', 'MID')  
quant

## mpg cylinders displacement horsepower weight acceleration year  
## MIN 9.00 3 68 46.0 1613.0 8.0 70  
## MAX 46.60 8 455 230.0 5140.0 24.8 82  
## MID 22.75 4 151 93.5 2803.5 15.5 76

1. What is the mean and standard deviation of each quantitative predictor?

quant <- **sapply**(auto[, 1**:**7], **function**(c) **c**(**mean**(c),**sd**(c)))   
**row.names**(quant) <- **c**('MEAN', 'STDDEV')

quant

## mpg cylinders displacement horsepower weight   
## MEAN 23.445918 5.471939 194.412 104.46939 2977.5842   
## STDDEV 7.805007 1.705783 104.644 38.49116 849.4026   
## acceleration year   
## MEAN 15.541327 75.979592  
## STDDEV 2.758864 3.683737

1. Remove the 25th through 115th observations. What is the range, median, mean, and standard deviation of each predictor in the subset of the data that remains?

quant <- sapply(auto[-(25:115), 1:7], function(c) c(range(c),median(c),mean(c),sd(c)))   
row.names(quant) <- c('MIN', 'MAX', 'MID', 'MEAN', 'STDDEV')

quant

## mpg cylinders displacement horsepower weight   
## MIN 11.000000 3.000000 68.00000 46.00000 1649.0000   
## MAX 46.600000 8.000000 455.00000 230.00000 4699.0000   
## MID 24.500000 4.000000 140.00000 90.00000 2720.0000   
## MEAN 25.015947 5.272425 180.46844 98.63455 2868.7708   
## STDDEV 7.637553 1.610032 95.90959 34.35597 755.6678   
## acceleration year  
## MIN 8.000000 70.00000  
## MAX 24.800000 82.00000  
## MID 15.600000 77.00000  
## MEAN 15.743189 77.20266  
## STDDEV 2.760072 3.31091

1. Suppose that we wish to predict gas mileage (mpg) on the basis of other variables. Using the full data set which variables do you believe will be useful in predicting mpg? Explain your answer using plots and correlation coefficients of the data.

res <- **sapply**(auto[, 2**:**7], **function**(c) **cor**(auto**$**mpg, c))  
res

## cylinders displacement horsepower weight acceleration year   
## -0.7776175 -0.8051269 -0.7784268 -0.8322442 0.4233285 0.5805410

full <- **plot**(auto[, 1**:**7])

Diagram, engineering drawing

Description automatically generated

# (Fig 1)

plt <-  
 **barplot**(res,  
 names.arg = **names**(res),  
 ylim = **c**(**-**1, 1),  
 ylab = 'Correlation Coefficients')  
**text**(  
 plt,  
 y = **round**(**unname**(res), 4),  
 label = **round**(**unname**(res), 4),  
 pos = 3,  
 col = "red"  
)

Chart, waterfall chart

Description automatically generated

# (Fig 2)

library(MASS)  
m <- lm(auto$mpg ~ ., auto[, 2:7])  
models <- stepAIC(m, direction = 'both', k = log(nrow(auto)))

## Start: AIC=1002.26  
## auto$mpg ~ cylinders + displacement + horsepower + weight + acceleration + year  
##   
## Df Sum of Sq RSS AIC  
## - horsepower 1 0.01 4543.4 996.29  
## - acceleration 1 8.24 4551.6 997.00  
## - cylinders 1 11.64 4555.0 997.29  
## - displacement 1 12.85 4556.2 997.40  
## <none> 4543.3 1002.26  
## - weight 1 1213.57 5756.9 1089.09  
## - year 1 2419.12 6962.5 1163.62  
##   
## Step: AIC=996.29  
## auto$mpg ~ cylinders + displacement + weight + acceleration +   
## year  
##   
## Df Sum of Sq RSS AIC  
## - cylinders 1 11.70 4555.1 991.33  
## - displacement 1 13.45 4556.8 991.48  
## - acceleration 1 14.36 4557.7 991.56  
## <none> 4543.4 996.29  
## + horsepower 1 0.01 4543.3 1002.26  
## - weight 1 1519.63 6063.0 1103.43  
## - year 1 2552.19 7095.5 1165.07  
##   
## Step: AIC=991.33  
## auto$mpg ~ displacement + weight + acceleration + year  
##   
## Df Sum of Sq RSS AIC  
## - displacement 1 3.45 4558.5 985.65  
## - acceleration 1 13.86 4568.9 986.55  
## <none> 4555.1 991.33  
## + cylinders 1 11.70 4543.4 996.29  
## + horsepower 1 0.06 4555.0 997.29  
## - weight 1 1552.07 6107.1 1100.30  
## - year 1 2549.22 7104.3 1159.58  
##   
## Step: AIC=985.65  
## auto$mpg ~ weight + acceleration + year  
##   
## Df Sum of Sq RSS AIC  
## - acceleration 1 10.5 4569.0 980.58  
## <none> 4558.5 985.65  
## + displacement 1 3.4 4555.1 991.33  
## + cylinders 1 1.7 4556.8 991.48  
## + horsepower 1 0.5 4558.0 991.58  
## - year 1 2594.4 7152.9 1156.29  
## - weight 1 9544.3 14102.8 1422.40  
##   
## Step: AIC=980.58  
## auto$mpg ~ weight + year  
##   
## Df Sum of Sq RSS AIC  
## <none> 4569.0 980.58  
## + acceleration 1 10.5 4558.5 985.65  
## + cylinders 1 5.0 4564.0 986.12  
## + horsepower 1 3.3 4565.7 986.27  
## + displacement 1 0.0 4568.9 986.55  
## - year 1 2752.3 7321.2 1159.43  
## - weight 1 11222.4 15791.3 1460.76

summary(models)

##  
## Call:  
## lm(formula = auto$mpg ~ weight + year, data = auto[, 2:7])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.8505 -2.3014 -0.1167 2.0367 14.3555   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.435e+01 4.007e+00 -3.581 0.000386 \*\*\*  
## weight -6.632e-03 2.146e-04 -30.911 < 2e-16 \*\*\*  
## year 7.573e-01 4.947e-02 15.308 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.427 on 389 degrees of freedom  
## Multiple R-squared: 0.8082, Adjusted R-squared: 0.8072   
## F-statistic: 819.5 on 2 and 389 DF, p-value: < 2.2e-16

models$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## auto$mpg ~ cylinders + displacement + horsepower + weight + acceleration +   
## year  
##   
## Final Model:  
## auto$mpg ~ weight + year  
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 385 4543.347 1002.2605  
## 2 horsepower 1 0.009440704 386 4543.356 996.2901  
## 3 cylinders 1 11.697280982 387 4555.054 991.3268  
## 4 displacement 1 3.447999525 388 4558.502 985.6521  
## 5 acceleration 1 10.450295633 389 4568.952 980.5785

confint(models)

## 2.5 % 97.5 %  
## (Intercept) -22.22439330 -6.470112738  
## weight -0.00705391 -0.006210241  
## year 0.66005097 0.854585595

plot(m)

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Through Fig 1, we can tell there exists some levels of correlations between MPG and cylinders, displacement, horsepower, weight, acceleration, and year. Specific correlation value shown in Fig 2. With running multilinear regression on all possible column combinations between MPG and cylinders, displacement, horsepower, weight, acceleration, and year, we can find that of all combinations, combining predictor weight and year yields the best Adjusted R-squared of 0.8072. Therefore, using weight and year will be useful in predicting MPG.