# FE590. Assignment #2.

### 2017-03-03

## Instructions

In this assignment, you should use R markdown to answer the questions below. Simply type your R code into embedded chunks as shown above. When you have completed the assignment, knit the document into a PDF file, and upload both the .pdf and .Rmd files to Canvas.

# Question 1 (based on JWHT Chapter 2, Problem 9)

Use the Auto data set from the textbook's website. When reading the data, use the options as.is = TRUE and na.strings="?". Remove the unavailable data using the na.omit() function.

```
library(ISLR)
setwd("/Users/apple/Desktop/590")
auto <- read.csv("Auto.csv", as.is = T, na.strings = "?")
auto <- na.omit(auto)
```

### 1. List the names of the variables in the data set.

```
colnames(auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

### 2. The columns origin and name are unimportant variables. Create a new data frame called cars that contains none of these unimportant variables

```
cars <- subset(auto, select = -c(origin, name))
```

### 3. What is the range of each quantitative variable? Answer this question using the range() function with the sapply() function (e.g., sapply(cars, range). Print a simple table of the ranges of the variables. The rows should correspond to the variables. The first column should be the lowest value of the corresponding variable, and the second column should be the maximum value of the variable. The columns should be suitably labeled.

```
car.range <- sapply(cars, range)
car.range <- t(car.range)
colnames(car.range) <- c("min", "max")
```

```
knitr::kable(car.range, caption = "The variable ranges of Cars")
```

Table 1: The variable ranges of Cars

| | min | max |
|---|---|---|
| mpg | 9 | 46.6 |
| cylinders | 3 | 8.0 |
| displacement | 68 | 455.0 |
| horsepower | 46 | 230.0 |
| weight | 1613 | 5140.0 |
| acceleration | 8 | 24.8 |
| year | 70 | 82.0 |

### 4. What is the mean and standard deviation of each variable? Create a simple table of the means and standard deviations.

```
vector.m <- sapply(cars, mean)
vector.sd <- sapply(cars, sd)
md <- rbind(vector.m, vector.sd)
md <- t(md)
colnames(md) <- c("mean", "standard_deviation")
```

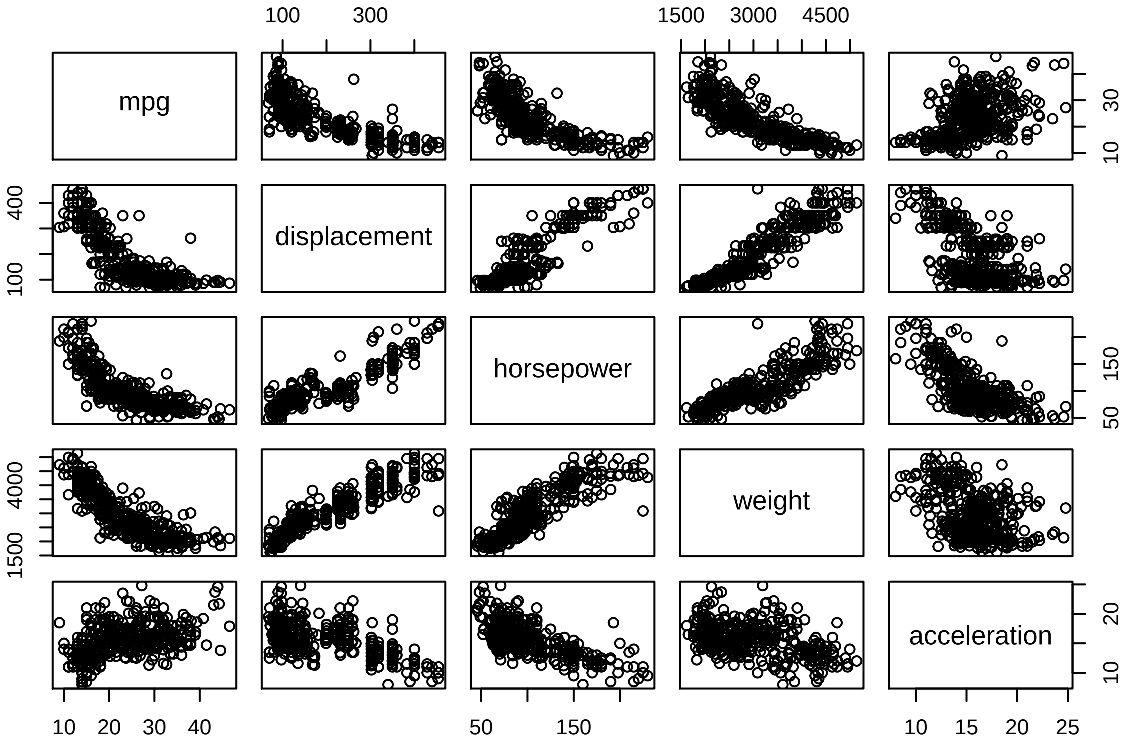
```
knitr::kable(md, caption = "Means and SD of Cars")
```

Table 2: Means and SD of Cars

| | mean | standard_deviation |
|---|---|---|
| mpg | 23.445918 | 7.805008 |
| cylinders | 5.471939 | 1.705783 |
| displacement | 194.411990 | 104.644004 |
| horsepower | 104.469388 | 38.491160 |
| weight | 2977.584184 | 849.402560 |
| acceleration | 15.541327 | 2.758864 |
| year | 75.979592 | 3.683737 |

### 5. Create a scatterplot matrix that includes the variables mpg, displacement, horsepower, weight, and acceleration using the pairs() function.

```
newdf <- subset(cars, select = c(mpg, displacement,horsepower, weight, acceleration))
pairs(newdf)
```




### 6. From the scatterplot, it should be clear that mpg has an almost linear relationship to predictors, and higher-order relationships to other variables. Using the regsubsets function in the leaps library, regress mpg onto

- displacement
- displacement squared
- horsepower
- horsepower squared
- weight
- weight squared
- acceleration

```
library(leaps)
attach(newdf)
newdf$displacement_squared <- displacement^2
newdf$horsepower_squared <- horsepower^2
newdf$weight_squared <- weight^2
detach(newdf)

m <- regsubsets(mpg~., data = newdf)
regs <- t(summary(m)$which)
```

Print a table showing what variables would be selected using best subset selection for all model orders.

```
knitr::kable(regs, caption = "Selections of variables")
```

Table 3: Selections of variables

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| (Intercept) | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |
| displacement | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE |
| horsepower | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE |
| weight | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE |
| acceleration | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE |
| displacement_squared | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE |
| horsepower_squared | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE |
| weight_squared | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE |

What is the most important variable affecting fuel consumption?

```
rownames(as.data.frame(coef(m, 1)))[2]
```

```
## [1] "weight"
```

What is the second most important variable affecting fuel consumption?

```
rownames(as.data.frame(coef(m, 2)))[3]
```

```
## [1] "weight_squared"
```

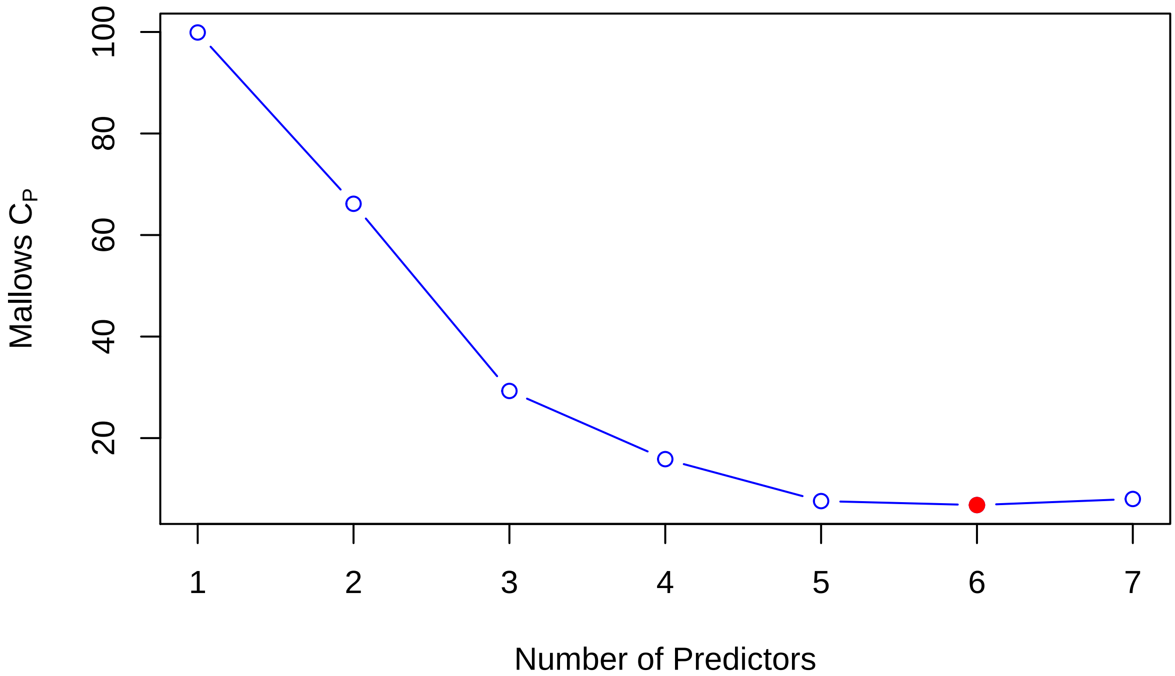
What is the third most important variable affecting fuel consumption?

```
rownames(as.data.frame(coef(m, 3)))[4]
```

```
## [1] "horsepower_squared"
```

## 7. Plot a graph showing Mallow's Cp as a function of the order of the model. Which model is the best?

```
cp=summary(m)$cp
i=which.min(cp)
plot(cp,type='b',col="blue",xlab="Number of Predictors",ylab=expression("Mallows C"[P]))
points(i,cp[i],pch=19,col="red")
```




The Mallow's Cp takes its minimum when the number of predictors is 6.

```
best_model <- summary(regsubsets(mpg~., data = newdf, nvmax = 6))$which
t(best_model[6,])
```

```
##      (Intercept) displacement horsepower weight acceleration
## [1,]        TRUE         TRUE       TRUE   TRUE         TRUE
##      displacement_squared horsepower_squared weight_squared
## [1,]                 TRUE               TRUE          FALSE
```

```
# The best model
```

## Question 2 (based on JWHT Chapter 3, Problem 10)

This exercise involves the Boston housing data set.

### 1. Load in the Boston data set, which is part of the MASS library in R. The data set is contained in the object Boston. Read about the data set using the command ?Boston. How many rows are in this data set? How many columns? What do the rows and columns represent?

```
library(MASS)
?Boston
```

There are 506 rows and 14 columns in the object Boston. Each row is an observation. For columns:

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: nitrogen 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 mean of 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.

black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat: lower status of the population (percent).

medv: median value of owner-occupied homes in $1000s.

## 2. Do any of the suburbs of Boston appear to have particularly high crime rates?

```r
summary(Boston)
```

```
##       crim                 zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

From the summry, we can see a wide range of crim rate (0.00632% - 88.97620%). However the mean and 3rd quantile is relatively low which indicate most parts of suburbs are save and other parts are very dangerous.

Tax rates?

The range of tax rate is 187% to 711% which is also quite large.

Pupil-teacher ratios?

Pupil-teacher ranges from 12.6% to 22% which is relatively low compared to former two indicators.

Comment on the range of each predictor.

### 3. How many of the suburbs in this data set bound the Charles river?

```
sum(Boston$chas)
```

```
## [1] 35
```

### 4. What is the median pupil-teacher ratio among the towns in this data set?

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

### 5. In this data set, how many of the suburbs average more than seven rooms per dwelling?

```
sum(Boston$rm > 7)
```

```
## [1] 64
```

More than eight rooms per dwelling?

```
sum(Boston$rm > 8)
```

```
## [1] 13
```

Comment on the suburbs that average more than eight rooms per dwelling.

```
summary(subset(Boston, rm > 8))
```

```
##       crim               zn            indus             chas
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000
##       nox               rm             age             dis
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907
##       rad              tax           ptratio          black
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9
```

```
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9
##      lstat            medv
##  Min.   :2.47   Min.   :21.9
##  1st Qu.:3.32   1st Qu.:41.7
##  Median :4.14   Median :48.3
##  Mean   :4.31   Mean   :44.2
##  3rd Qu.:5.12   3rd Qu.:50.0
##  Max.   :7.44   Max.   :50.0
```

Compared to other suburbs, these suburbs that average more than eight rooms per dwelling have more lower status of the population (lstat).

# Question 3 (based on JWHT Chapter 4, Problem 10)

This question should be answered using the Weekly data set, which is part of the ISLR package. This data contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

## 1. What does the data represent?

```
library(ISLR)
?Weekly
```

The data represents weekly percentage returns for the S&P 500 stock index between 1990 and 2010.

Year: The year that the observation was recorded

Lag1: Percentage return for previous week

Lag2: Percentage return for 2 weeks previous

Lag3: Percentage return for 3 weeks previous

Lag4: Percentage return for 4 weeks previous

Lag5: Percentage return for 5 weeks previous

Volume: Volume of shares traded (average number of daily shares traded in billions)

Today: Percentage return for this week

Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week

## 2. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
attach(Weekly)
glm.fit <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, family=binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

```r
detach(Weekly)
```

The predictor "Lag2" is statistically significant.

### 3. Fit a logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

```r
Weekly.train <- subset(Weekly, Year <= 2008)
Weekly.test <- subset(Weekly, Year > 2008)
fit <- glm(Direction ~ Lag2, data = Weekly.train, family = binomial)
glm.probs <- predict(fit, Weekly.test, type ="response")

glm.pred <- rep("Down", nrow(Weekly.test))
glm.pred[glm.probs >.5] <- "Up"

table(glm.pred, Weekly.test$Direction)
```

```
##
## glm.pred Down Up
##     Down    9  5
##     Up     34 56
```

```r
glm.ratio <- mean(glm.pred == Weekly.test$Direction)
glm.ratio
```

```
## [1] 0.625
```

The fraction of correct predictions is 0.625.

## 4. Repeat Part 3 using LDA.

```r
lda.fit <- lda(Direction ~ Lag2, data = Weekly.train)
lda.pred <- predict(lda.fit, Weekly.test)
lda.class <- lda.pred$class


table(lda.class, Weekly.test$Direction)
```

```
##
## lda.class Down Up
##      Down    9  5
##      Up     34 56
```

```r
lda.ratio <- mean(lda.class == Weekly.test$Direction)
lda.ratio
```

```
## [1] 0.625
```

The fraction of correct predictions is 0.625.

## 5. Repeat Part 3 using QDA.

```r
qda.fit <- qda(Direction ~ Lag2, data = Weekly.train)
qda.pred <- predict(qda.fit, Weekly.test)
qda.class <- qda.pred$class

table(qda.class, Weekly.test$Direction)
```

```
##
## qda.class Down Up
##      Down    0  0
##      Up     43 61
```

```r
qda.ratio <- mean(qda.class == Weekly.test$Direction)
qda.ratio
```

```
## [1] 0.5865385
```

The fraction of correct predictions is 0.587.

## 6. Repeat Part 3 using KNN with K = 1, 2, 3.

```r
library (class)
direction <- Weekly.train$Direction
Train.knn <- as.matrix(Weekly.train[, 3])
```

```r
Test.knn <- as.matrix(Weekly.test[, 3])

knn.pred1 <- knn(Train.knn, Test.knn, direction, k = 1)
table(knn.pred1, Weekly.test$Direction)
```

```
##
## knn.pred1 Down Up
##      Down   21 29
##      Up     22 32
```

```r
knn1.ratio <- mean(knn.pred1 == Weekly.test$Direction)
knn1.ratio
```

```
## [1] 0.5096154
```

```r
knn.pred2 <- knn(Train.knn, Test.knn, direction, k = 2)
table(knn.pred2, Weekly.test$Direction)
```

```
##
## knn.pred2 Down Up
##      Down   21 29
##      Up     22 32
```

```r
knn2.ratio <- mean(knn.pred2 == Weekly.test$Direction)
knn2.ratio
```

```
## [1] 0.5096154
```

```r
knn.pred3 <- knn(Train.knn, Test.knn, direction, k = 3)
table(knn.pred3, Weekly.test$Direction)
```

```
##
## knn.pred3 Down Up
##      Down   16 19
##      Up     27 42
```

```r
knn3.ratio <- mean(knn.pred3 == Weekly.test$Direction)
knn3.ratio
```

```
## [1] 0.5576923
```

## 7. Which of these methods in Parts 3, 4, 5, and 6 appears to provide the best results on this data?

```r
method <- data.frame(glm.ratio, lda.ratio, qda.ratio, knn1.ratio, knn2.ratio, knn3.ratio)
```

```r
knitr::kable(method)
```

| glm.ratio | lda.ratio | qda.ratio | knn1.ratio | knn2.ratio | knn3.ratio |
|---:|---:|---:|---:|---:|---:|
| 0.625 | 0.625 | 0.5865385 | 0.5096154 | 0.5096154 | 0.5576923 |

The logistic regression model and LDA model provide the best results on this data.

## Question 4

**Write a function that works in R to gives you the parameters from a linear regression on a data set between two sets of values (in other words you only have to do the 2-D case). Include in the output the standard error of your variables. You cannot use the lm command in this function or any of the other built in regression models. For example your output could be a 2x2 matrix with the parameters in the first column and the standard errors in the second column. For up to 5 bonus points, format your output so that it displays and operates similar in function to the output of the lm command.(i.e. in a data frame that includes all potentially useful outputs)**

```
linear.regression <- function(data1, data2) {
    m1 <- mean(data1) # means of data
    m2 <- mean(data2)
    n <- length(data1)
    b <- (sum(data1*data2) - n*m1*m2)/(sum(data1^2) - n*m1^2)
    a <- m2 - b*m1
    data2.hat <- a + b*data1
    epsilon <- data2.hat - data2
    se.b <- sqrt(n*sum(epsilon^2)/(n-2)/(n*sum(data1^2)-(sum(data1))^2))
    se.a <- se.b*sqrt(sum(data1^2)/n)
    t.b <- b/se.b
    t.a <- a/se.a
    pr.b <- dt(t.b, n-2)
    pr.a <- dt(t.a, n-2)

    Estimate <- c(a, b)
    Std.Error <- c(se.a, se.b)
    t_value <- c(t.a, t.b)
    Pr <- c(pr.a, pr.b)
    df <- data.frame(Estimate, Std.Error, t_value, Pr)
    rownames(df) <- c("(Intercept)","slope")
    return(df)
}
linear.regression(Weekly$Lag1, Weekly$Lag2)
```

```
##                Estimate  Std.Error   t_value         Pr
## (Intercept)  0.16235187 0.07140973  2.273526 0.03020116
## slope       -0.07486073 0.03024891 -2.474824 0.01876503
```

**Compare the output of your function to that of the lm command in R.**

```
LG <- lm(Lag1 ~ Lag2, data = Weekly)
summary(LG)
```

```
##
## Call:
## lm(formula = Lag1 ~ Lag2, data = Weekly)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -19.0604  -1.2715   0.1134   1.2796  11.2362
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.16189    0.07140   2.267   0.0236 *
## Lag2        -0.07485    0.03024  -2.475   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.351 on 1087 degrees of freedom
## Multiple R-squared:  0.005603,   Adjusted R-squared:  0.004688
## F-statistic: 6.125 on 1 and 1087 DF,  p-value: 0.01348
```