Lab 1

Nancy Zhou (nz2ay) and Philip Lee (phl6cw)

Note: We obtained permission from Professor Woo to work in a group of 2, as there weren't enough people to form groups of 3-4.

Problem 1: How many observations are there in this data set? Using the dim() function will be useful.

```
data=Auto[0:8] #drops 'name' variable
dim(data)
```

```
## [1] 392   8
```

There are 392 observations with 8 variables

Problem 2: How many variables are there in this data set? How many of the variables are quantitative? How many of the variables are categorical? Typing ?Auto to read the documentation will be useful.

```
#?Auto
year=factor(data$year) #convert to categorical
origin=factor(data$origin) #convert to categorical
is.factor(year)
```

```
## [1] TRUE
```

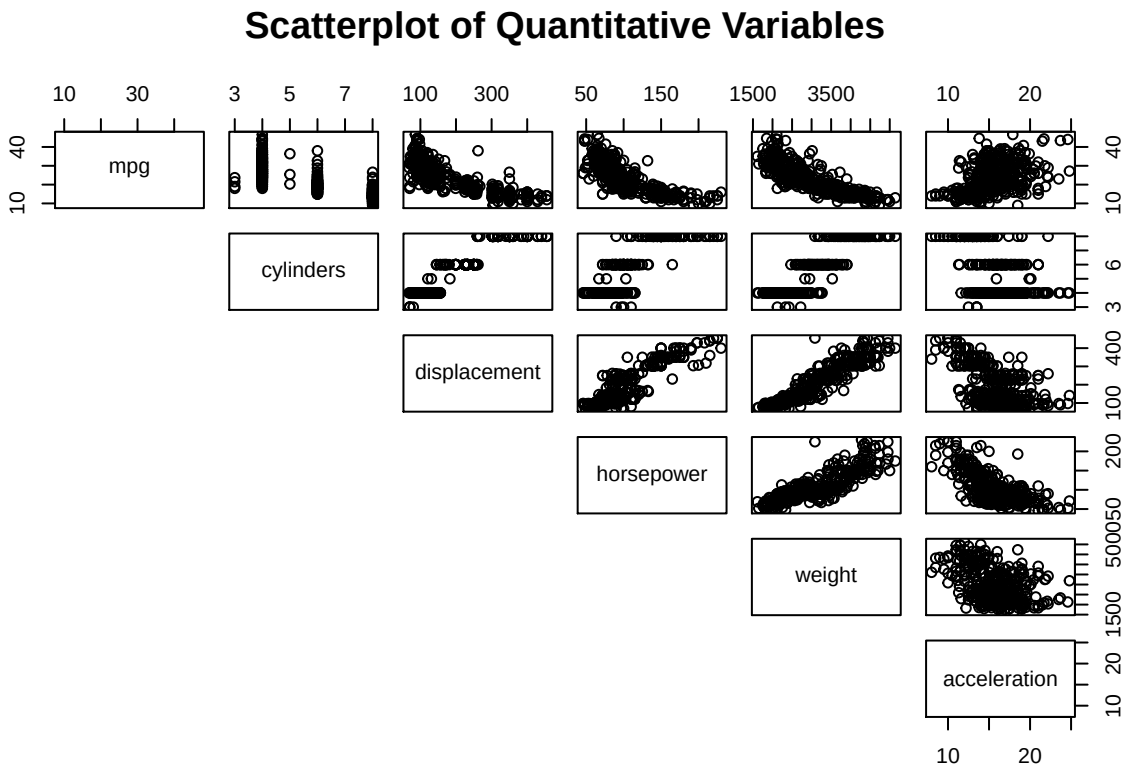
```
is.factor(origin)
```

```
## [1] TRUE
```

By typing ?Auto, we are able to see that mpg, number of cylinders, engine displacement (inches), horsepower, weight, acceleration were quantiative variables while model year and country of origin were categorical/qualitative variables. We found that there are 6 quantitative variables and 2 qualitative variables with 8 total variables in the dataset.

Problem 3: Produce a scatterplot matrix for all the quantitative variables in the data set.

```
pairs(data[,0:6], lower.panel = NULL, main="Scatterplot of Quantitative Variables")
```



Problem 4: Produce a correlation matrix for all the quantitative variables in the data set.

```
round(cor(data[,0:6]),3)
```

```
##                mpg cylinders displacement horsepower weight acceleration
## mpg          1.000    -0.778       -0.805     -0.778 -0.832        0.423
## cylinders   -0.778     1.000        0.951      0.843  0.898       -0.505
## displacement -0.805     0.951        1.000      0.897  0.933       -0.544
## horsepower  -0.778     0.843        0.897      1.000  0.865       -0.689
## weight      -0.832     0.898        0.933      0.865  1.000       -0.417
## acceleration  0.423    -0.505       -0.544     -0.689 -0.417        1.000
```

Problem 5: What relationships do you see, if any, among the variables based on your output from questions 3 and 4?

We see that engine displacement, the number of clynders and vehicle weight are all very positively correlated with each other. Miles per gallon, on the otherhand, seems strongly negatively correlated with all other variables with the exception of acceleration, with which it has a weaker relationship. Additionally, scatterplots related to number of clynders are evenly

distributed into columns or rows, suggesting that data was collected in a regimented fashion (as you could only have between 4 to 8 cylinders for each observation).

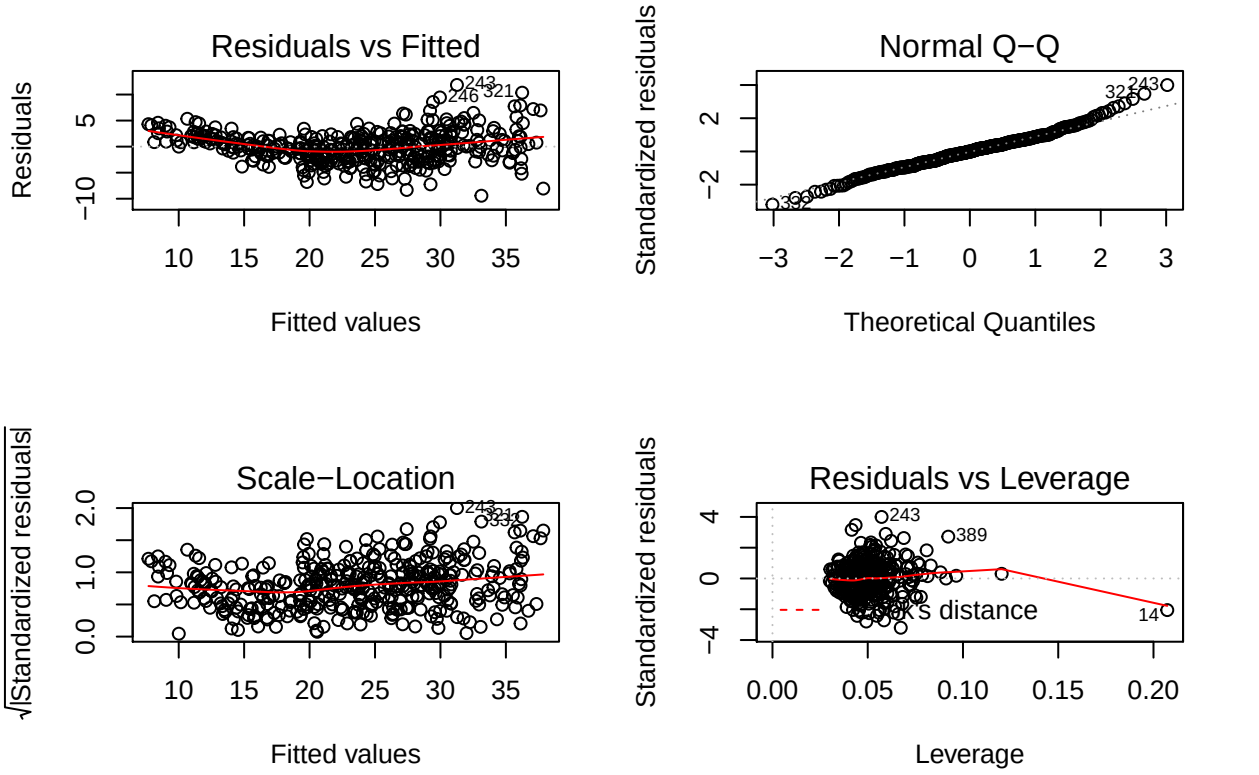
Problem 7: Fit a multiple linear regression model with mpg as the response variable and all the other variables (except name) as predictors. Create the necessary plots to assess if the model assumptions are met. Are there transformations you should try?

```
result<-lm(data$mpg~data$cylinders+data$displacement
           +data$horsepower+data$weight+data$acceleration+year+origin)
summary(result)
```

```
##
## Call:
## lm(formula = data$mpg ~ data$cylinders + data$displacement +
##     data$horsepower + data$weight + data$acceleration + year +
##     origin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4288 -1.9194 -0.0287  1.7899 11.8399
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       37.0199278  2.1417790  17.285  < 2e-16 ***
## data$cylinders    -0.1924289  0.3040482  -0.633 0.527195
## data$displacement  0.0172507  0.0072044   2.394 0.017139 *
## data$horsepower   -0.0240199  0.0136328  -1.762 0.078904 .
## data$weight       -0.0061203  0.0006494  -9.424  < 2e-16 ***
## data$acceleration  0.0543880  0.0919344   0.592 0.554480
## year71             1.0461869  0.8730131   1.198 0.231538
## year72             0.0330325  0.8531036   0.039 0.969134
## year73            -0.5322929  0.7718143  -0.690 0.490835
## year74             1.6545531  0.9129699   1.812 0.070750 .
## year75             0.9415172  0.8953739   1.052 0.293695
## year76             1.7486166  0.8573480   2.040 0.042100 *
## year77             3.2399161  0.8759807   3.699 0.000249 ***
## year78             3.0821303  0.8333179   3.699 0.000249 ***
## year79             5.3812526  0.8791655   6.121 2.36e-09 ***
## year80             9.5116004  0.9339482  10.184  < 2e-16 ***
## year81             6.9070845  0.9223997   7.488 5.12e-13 ***
## year82             8.6173419  0.9031369   9.542  < 2e-16 ***
## origin2            2.5075851  0.5316558   4.717 3.40e-06 ***
## origin3            2.5002584  0.5225230   4.785 2.47e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.05 on 372 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8473
## F-statistic: 115.2 on 19 and 372 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(result)
```



The residual plot with the red line indicates some curvature, indicating that the form of our model may not be reasonable.The QQ plot suggests that the residuals are normal. The scale-location plot indicates that variance is not constant as the vertical spread of plot is not constant. Lastly, the residuals vs. leverage plot does not indicate any influential outliers based on Cook's distance line.

Since the variance is not constant, we may consider tranforming the response variable, miles per gallon in such a manner that the resulting transformed variable meets the assumptions of the analysis. Once this is accomplished, one can perform the analysis on the transformed variable.