Simple & Multiple Linear Regression Analysis

```
## [1] 398   9

##   V1 V2  V3    V4   V5   V6 V7 V8                          V9
## 1 18  8 307 130.0 3504 12.0 70  1 chevrolet chevelle malibu
## 2 15  8 350 165.0 3693 11.5 70  1         buick skylark 320
## 3 18  8 318 150.0 3436 11.0 70  1        plymouth satellite
## 4 16  8 304 150.0 3433 12.0 70  1             amc rebel sst
## 5 17  8 302 140.0 3449 10.5 70  1               ford torino
## 6 15  8 429 198.0 4341 10.0 70  1          ford galaxie 500

## [1] "numeric"

## [1] "integer"

## [1] "numeric"

## [1] "character"

## [1] "numeric"

## [1] "numeric"

## [1] "integer"

## [1] "integer"

## [1] "character"

## Warning: NAs introduced by coercion

## [1] 398   8

## [1] "numeric"

## [1] 392   8
```

PART 1: SIMPLE LINEAR REGRESSION ANALYSIS

This part aims at examining the association between the response variable and a specific predictor variable you choose. In your report, make sure you

(1) Provide the justification of determining the response Y and predictor X variables.

Answer:

List of variables:

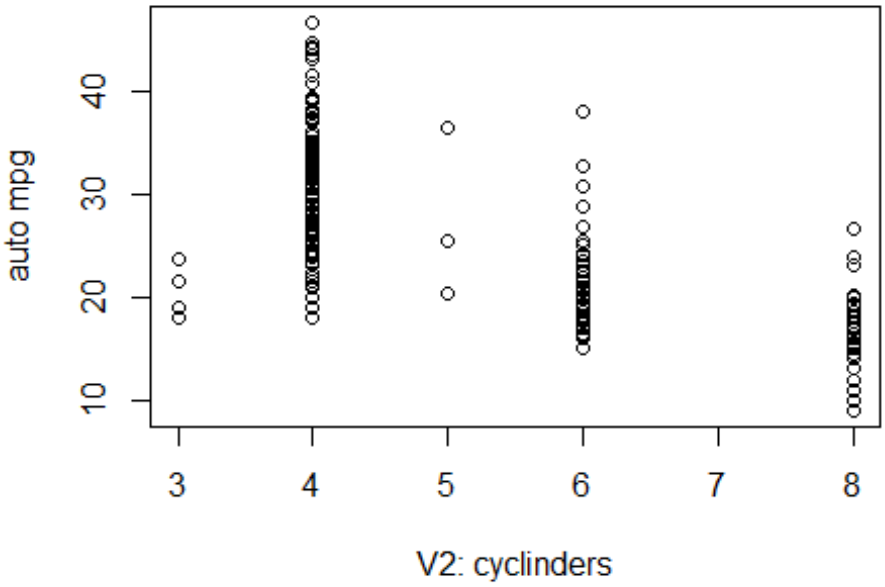
1. mpg: continuous

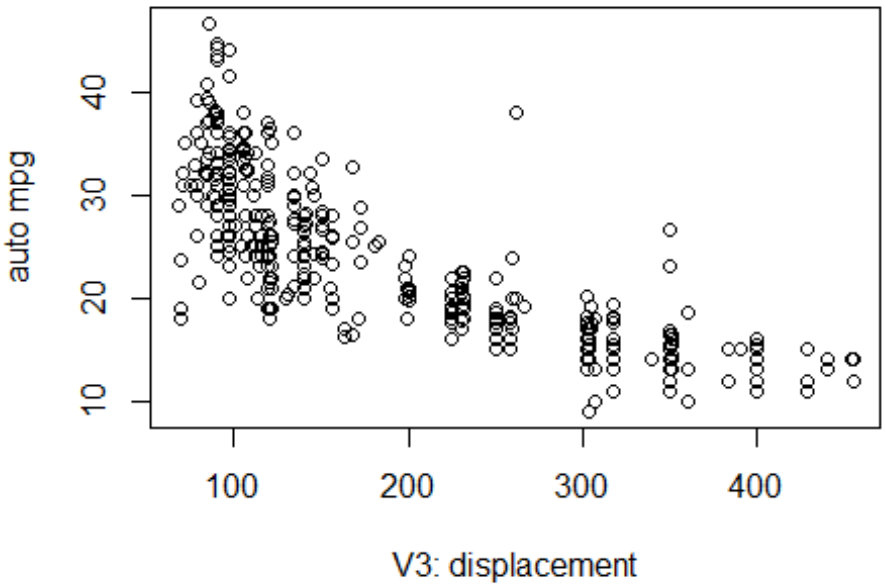
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

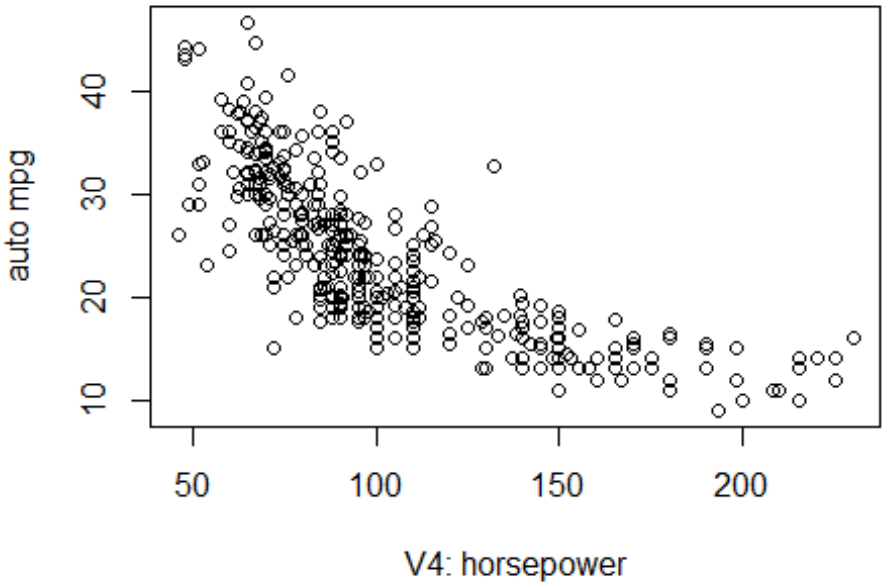
We chose the Auto MPG dataset. The dataset contains 398 observations with 9 variables (6 continuous, 3 discrete). The variables are listed above. Before conducting the analysis, had a data cleansing process of: 1)dropping the variable V9: car names, 2) changing the class of V4 variable from character to numeric, and 3)deleting rows with missing values (6 missing values were deleted, resulting in total of 392 observations).

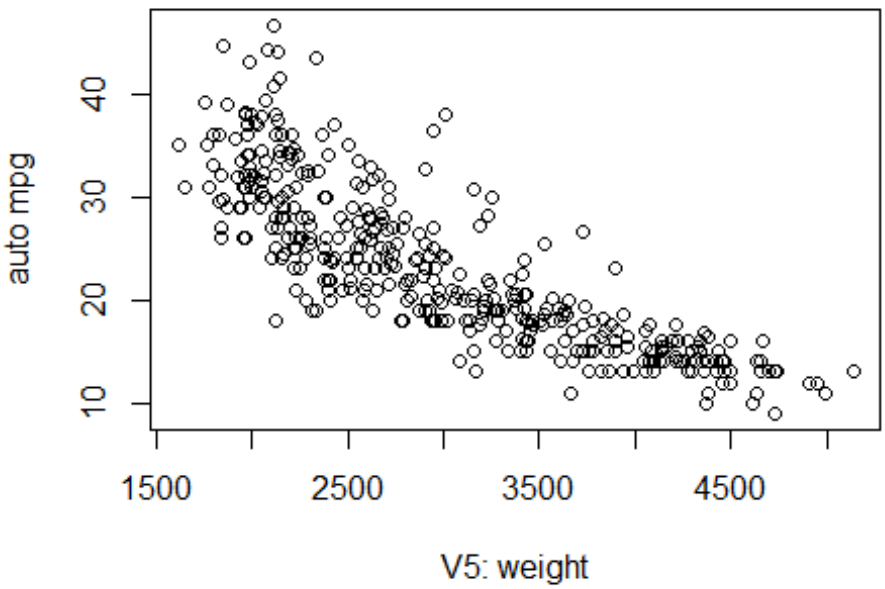
We would say that the fuel efficiency of a car is the most important factor that one would be interested in, and in this respect we selected MPG (Mileage per Gallon) as our response variable (Y).

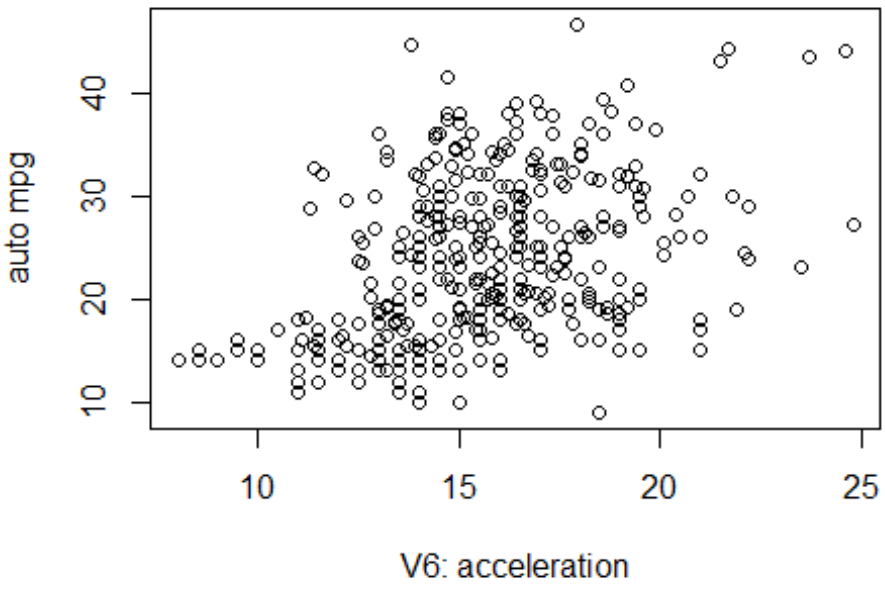
For determining the predictor variable, X, we first examined scatterplots of V2 ~ V8 to V1. According to the R results, V3: displacement, V4: horsepower, and V5: weight seemed to show the strongest relationship with the response variable, V1: auto MPG (See the below scatterplots). These three variables showed negative relationship with the response variable. Next, we examined the R square values when regressing each variable to auto mpg. (see the below results) We chose V5: Weight as the predictor variable because the variable showed the highest R square value of 0.6926.

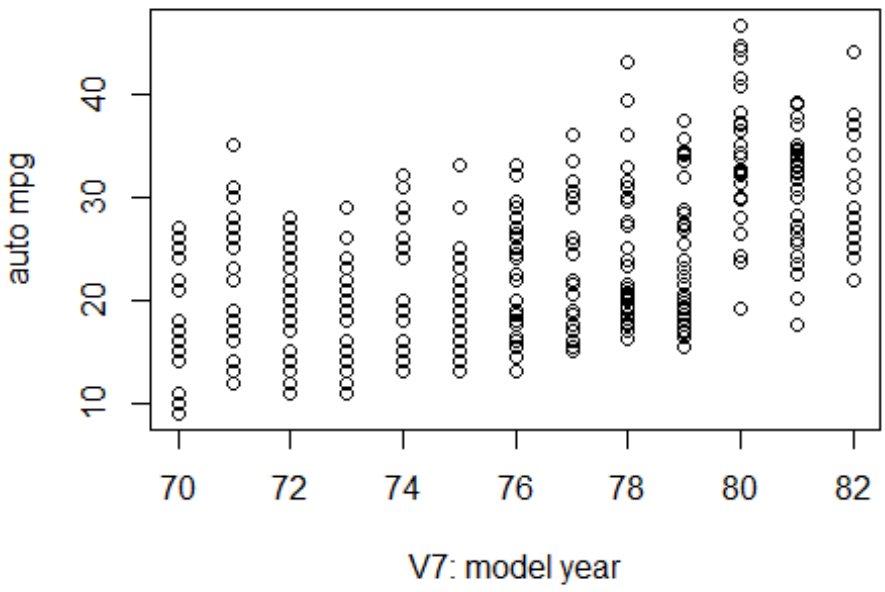

auto mpg
10
20
30
40
3
4
5
6
7
8
V2: cyclinders

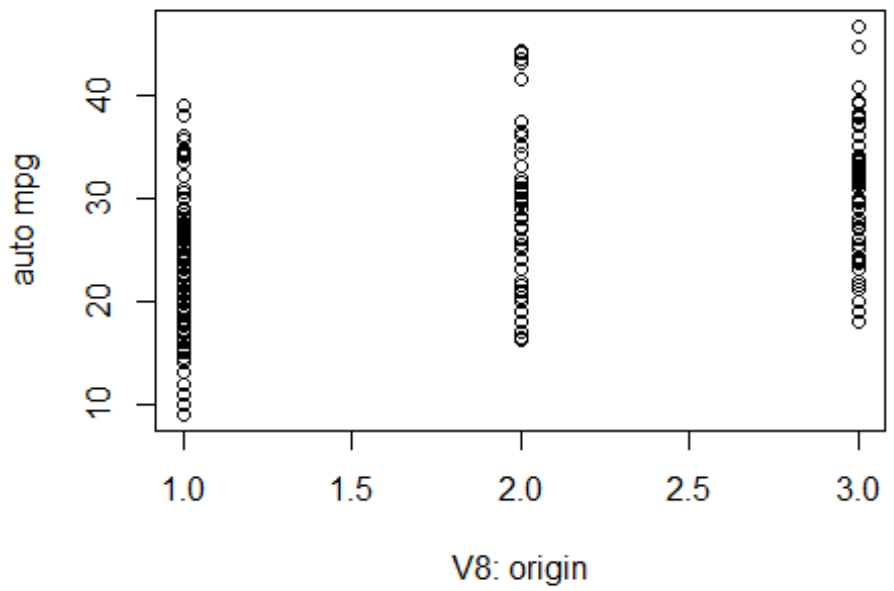

auto mpg
10
20
30
40
100
200
300
400
V3: displacement


auto mpg
10
20
30
40
50
100
150
200
V4: horsepower


auto mpg
10
20
30
40
1500
2500
3500
4500
V5: weight


auto mpg
10
20
30
40
10
15
20
25
V6: acceleration


auto mpg
10
20
30
40
70
72
74
76
78
80
82
V7: model year




```
##
## Call:
## lm(formula = V1 ~ V2, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2413  -3.1832  -0.6332   2.5491  17.9168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.9155     0.8349   51.40   <2e-16 ***
## V2           -3.5581     0.1457  -24.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.914 on 390 degrees of freedom
## Multiple R-squared:  0.6047, Adjusted R-squared:  0.6037
## F-statistic: 596.6 on 1 and 390 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = V1 ~ V3, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9170  -3.0243  -0.5021   2.3512  18.6128
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.12064    0.49443   71.03   <2e-16 ***
## V3          -0.06005    0.00224  -26.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.635 on 390 degrees of freedom
## Multiple R-squared:  0.6482, Adjusted R-squared:  0.6473
## F-statistic: 718.7 on 1 and 390 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = V1 ~ V4, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## V4          -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = V1 ~ V5, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9736  -2.7556  -0.3358   2.1379  16.5194
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.216524   0.798673   57.87   <2e-16 ***
## V5          -0.007647   0.000258  -29.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.333 on 390 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6918
## F-statistic: 878.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = V1 ~ V6, data = autompg1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.989  -5.616  -1.199   4.801  23.239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8332     2.0485   2.359   0.0188 *
## V6            1.1976     0.1298   9.228   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.08 on 390 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1771
## F-statistic: 85.15 on 1 and 390 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = V1 ~ V7, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0212  -5.4411  -0.4412   4.9739  18.2088
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -70.01167    6.64516  -10.54   <2e-16 ***
## V7            1.23004    0.08736   14.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.363 on 390 degrees of freedom
## Multiple R-squared:  0.337,  Adjusted R-squared:  0.3353
## F-statistic: 198.3 on 1 and 390 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = V1 ~ V8, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2416  -5.2533  -0.7651   3.8967  18.7115
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.8120     0.7164   20.68   <2e-16 ***
## V8            5.4765     0.4048   13.53   <2e-16 ***
```

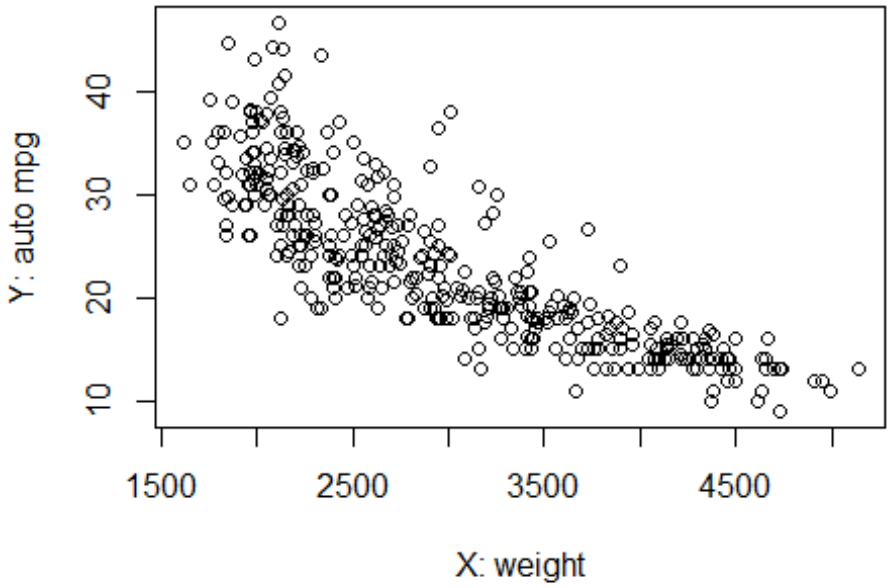
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.447 on 390 degrees of freedom
## Multiple R-squared:  0.3195, Adjusted R-squared:  0.3177
## F-statistic: 183.1 on 1 and 390 DF,  p-value: < 2.2e-16
```

(2) Construct a scatterplot of Y against X. Comment on the association between the two variables.

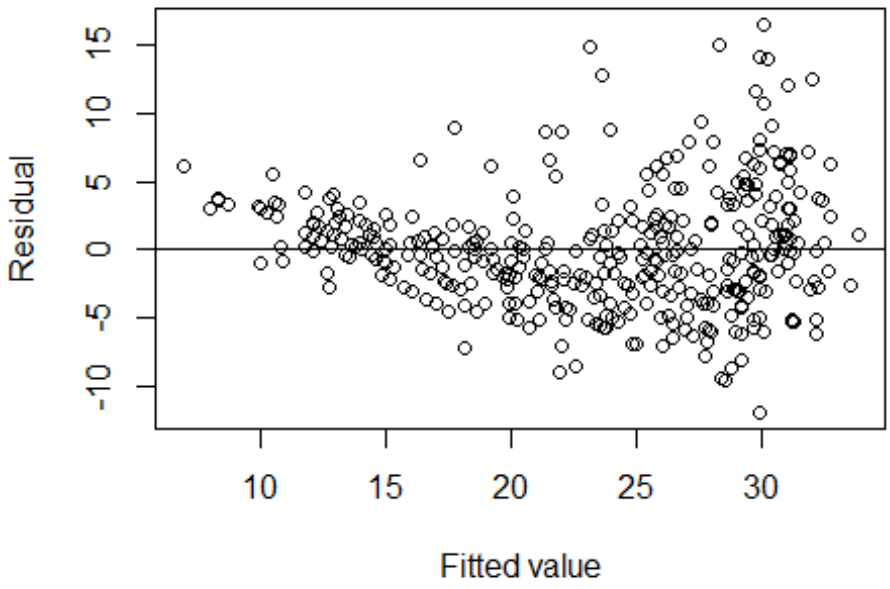
- If the association is not linear, apply transformations on either Y or X, or on both, to make a linear model more appropriate for the transformed data.
- When transformations are used, re-construct a scatterplot of (transformed) Y against (transformed) X and work on the transformed data for parts (3)-(7).

Answer:

See the following scatterplot. X and Y show a negative relation, but it seems nonlinear so need for transformations are suggested.

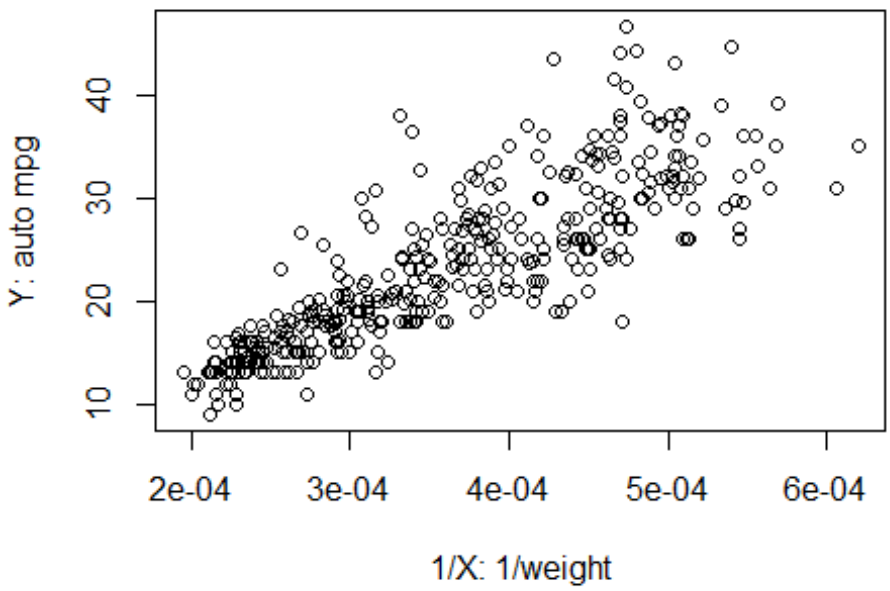


We also examined a graph of e on the fitted values(see the below graph), and the graph is somewhat different from a random scattering. Thus, violations on the linear relationship between X and Y is suggested.

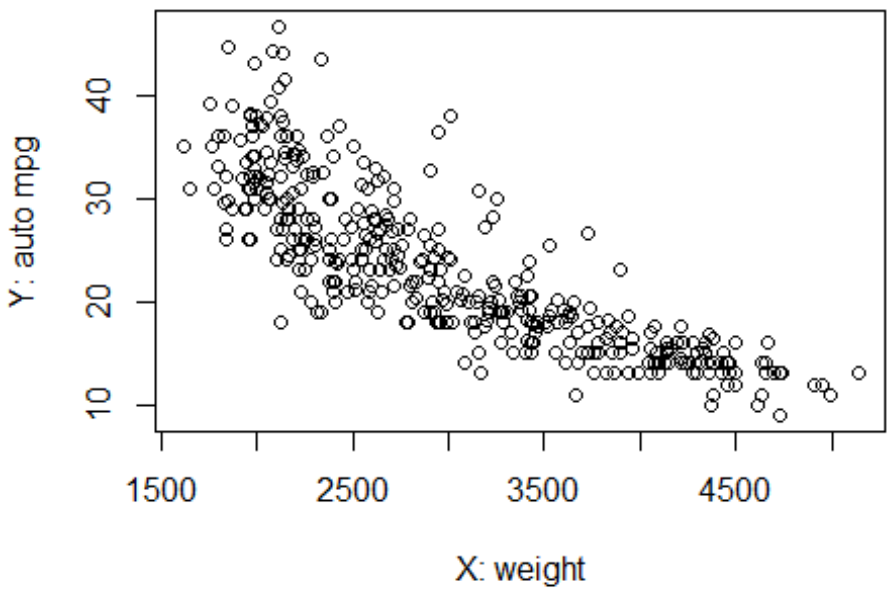


Based on the graphical results, we concluded that some transformations on X and/or Y are suggested. Referring to the prototype learned from the lecture, we tried several transformations on X (exponential, inverse) and Y (log, square root, inverse). Based on the scatterplots after trnasformation, we concluded that log transformation on Y is most appropriate.Thus, we hereafter use the log of Y as the response variable in Part 1. We denote log Y as Y'.

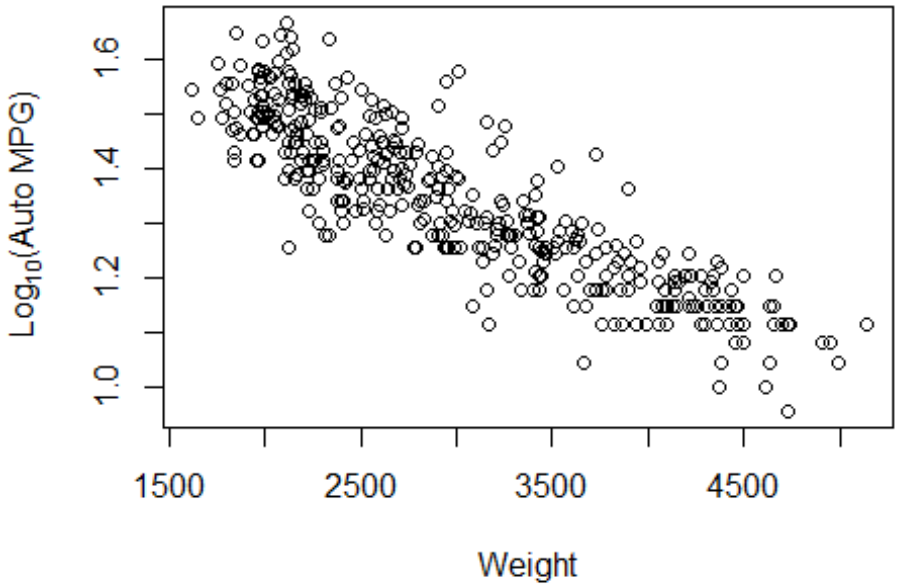
```
# Transformations on X:
V5.exp=exp(autompg1$V5); V5.inv=1/autompg1$V5;

autompg1$V5.exp <- V5.exp
autompg1$V5.inv <- V5.inv


names(autompg1)

##  [1] "V1"     "V2"     "V3"     "V4"     "V5"     "V6"     "V7"     "V8"
##  [9] "V5.exp" "V5.inv"

head(autompg1$V5.exp)

## [1] Inf Inf Inf Inf Inf Inf

#plot(V1~V5.exp,data=autompg1,xlab="exp(X): exp(weight)",ylab="Y: auto mpg")
#not appropriate
plot(V1~V5.inv,data=autompg1,xlab="1/X: 1/weight",ylab="Y: auto mpg")
```



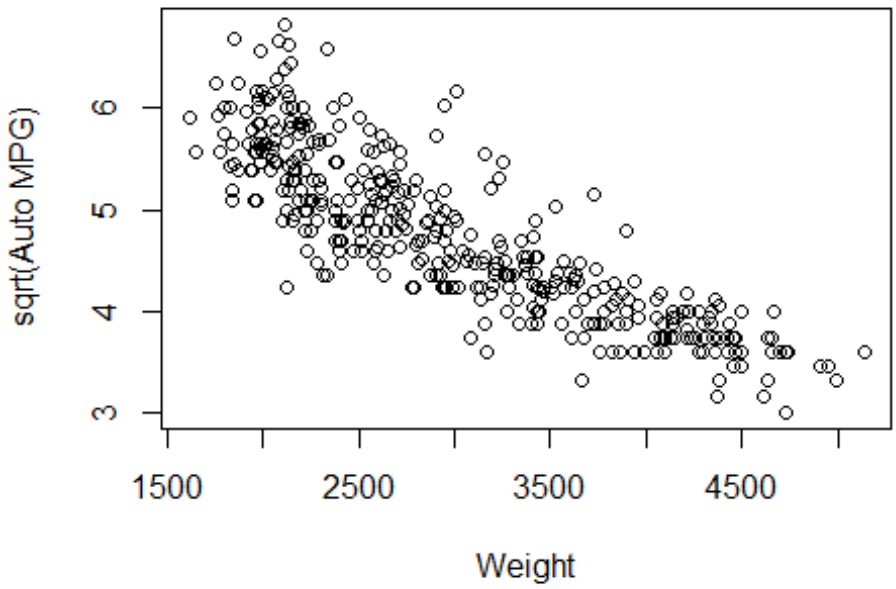

```
plot(V1~V5,data=autompg1,xlab="X: weight",ylab="Y: auto mpg")
```



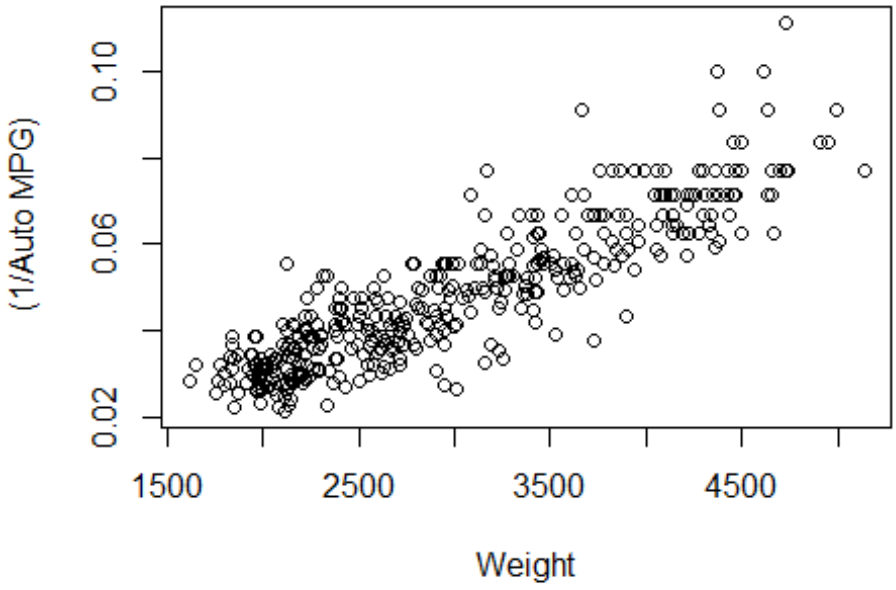
```
# Transformations on Y:
V1.log=log10(autompg1$V1); V1.sqrt=sqrt(autompg1$V1);V1.inv=1/(autompg1$V1);
autompg1$V1.log <- V1.log
autompg1$V1.sqrt <- V1.sqrt
autompg1$V1.inv <- V1.inv

plot(V1.log~V5,data=autompg1,xlab="Weight",ylab=expression(paste(Log[10],"(Au
to MPG)")))
```




```
plot(V1.sqrt~V5,data=autompg1,xlab="Weight",ylab=expression(paste(sqrt,"(Auto
MPG)")))
```



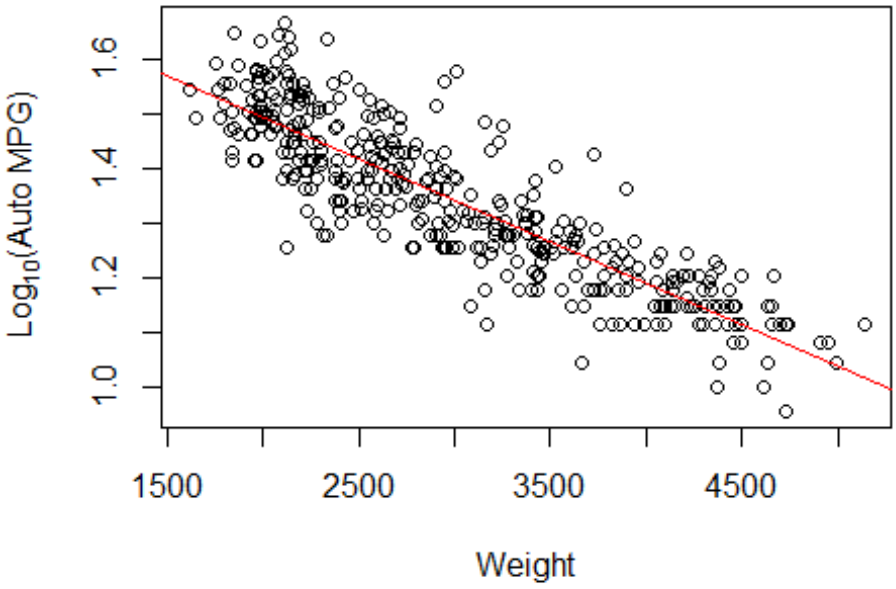

```
plot(V1.inv~V5,data=autompg1,xlab="Weight",ylab=expression(paste("(1/Auto
MPG)")))
```



(3) Regress (transformed) Y on (transformed) X. State the estimated regression function and

superimpose it on the scatterplot.

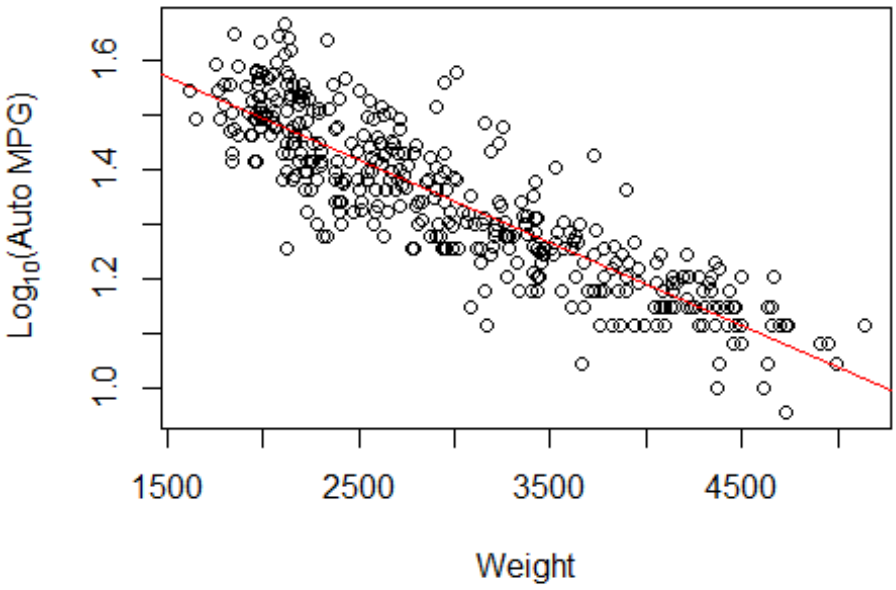
```
##
## Call:
## lm(formula = V1.log ~ V5, data = autompg1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.220259 -0.043281 -0.002697  0.043314  0.239900
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.799e+00  1.316e-02  136.66   <2e-16 ***
## V5          -1.522e-04  4.252e-06  -35.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07141 on 390 degrees of freedom
## Multiple R-squared:  0.7668, Adjusted R-squared:  0.7662
## F-statistic:  1282 on 1 and 390 DF,  p-value: < 2.2e-16
```



Answer: Regression function:

$$Y' = 1.799 - 0.0001522X + \epsilon$$

(4) Report the values of MSE and R^2. Interpret them, respectively, in the context of the problem.




```
##
## Call:
## lm(formula = V1.log ~ V5, data = autompg1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.220259 -0.043281 -0.002697  0.043314  0.239900
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.799e+00  1.316e-02  136.66   <2e-16 ***
## V5          -1.522e-04  4.252e-06  -35.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07141 on 390 degrees of freedom
## Multiple R-squared:  0.7668, Adjusted R-squared:  0.7662
## F-statistic:  1282 on 1 and 390 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Response: V1.log
##            Df Sum Sq Mean Sq F value    Pr(>F)
## V5          1 6.5384  6.5384  1282.2 < 2.2e-16 ***
```

```
## Residuals 390 1.9887  0.0051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 1.98872
```

```
## [1] 0.005099282
```

Answer:

Based on the output, the coefficient of determination is R^2= 0.7668, which means 76.68% of the variation in log10(Mileage per Gallon) can be accounted by Weight of the car into the regression model. The value of Mean Squared Error(MSE) is 0.005099282. The MSE that we get is pretty small, which means we get the good estimator. The above data is scattered wildly around the regression line, so 0.005099282 is as good as it gets.

(5) Test whether $\beta 1$ is different from zero or not at 0.05 level of significance. State the alternatives, the value of test statistic, p-value, and your conclusion.

```
summary(m1)
```

```
##
## Call:
## lm(formula = V1.log ~ V5, data = autompg1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.220259 -0.043281 -0.002697  0.043314  0.239900
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.799e+00  1.316e-02  136.66   <2e-16 ***
## V5          -1.522e-04  4.252e-06  -35.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07141 on 390 degrees of freedom
## Multiple R-squared:  0.7668, Adjusted R-squared:  0.7662
## F-statistic:  1282 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
n <- nrow(autompg1)
n
```

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

```
qt(0.975,n-2) ## t-crit value
```

```
## [1] 1.966065
```

```
tstat <- -1.522e-04/4.252e-06 ## t-stat
tstat
```

```
## [1] -35.79492
```

Answer:

(1) Alternatives: H_0: β 1 = 0 H_a: β 1 $\neq$ 0

(2) Decision rule:

Critical value method: if |t*| is less than or equal to the t-critical value, 1.966065, conclude Ho, if abs(t-stat) > t-critical value, 1.966065, then conclude Ha.

P-value Method : if the p-value of the test is less than alpha = 0.05, conclude Ha.

(3) The Value of test statistic, p-value and Conclusion

t-stat is -35.79492, in which the absolute value is larger than the t critical value, 1.966065, so we conclude Ha and conclude that there is a linear association between X and log Y.

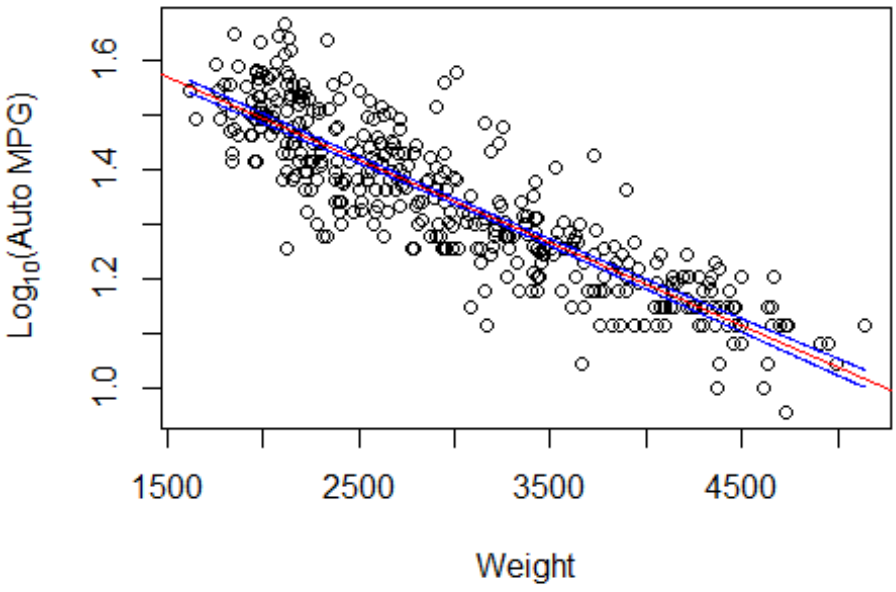
p-value is 2.2e-16, which is smaller than the alpha = 0.05 so we conclude Ha.

In conclusion, we can conclude that there exists a linear relationship between the log of auto MPG(V1) and weight (V5).

(6) Superimpose the (pointwise) 90% confidence band on the scatterplot.

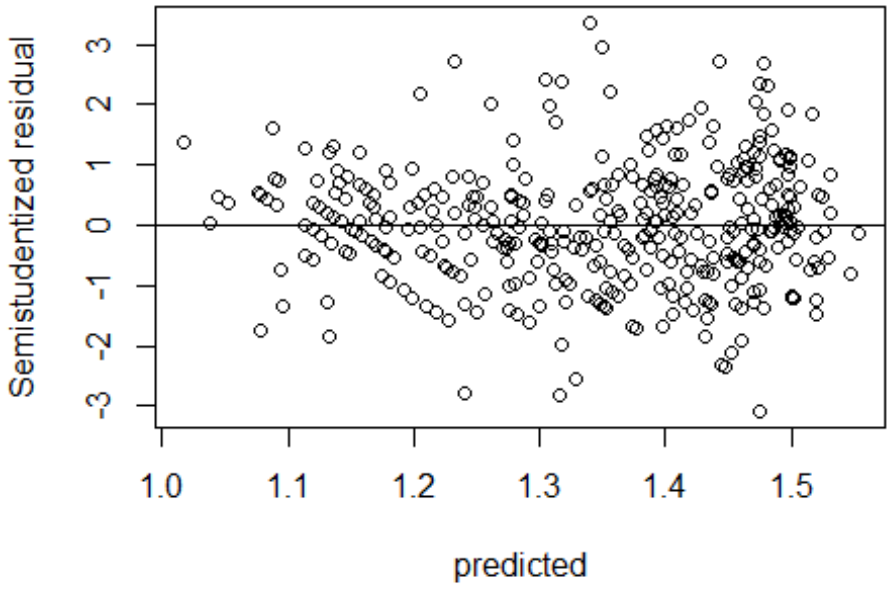
```
## integer(0)
```

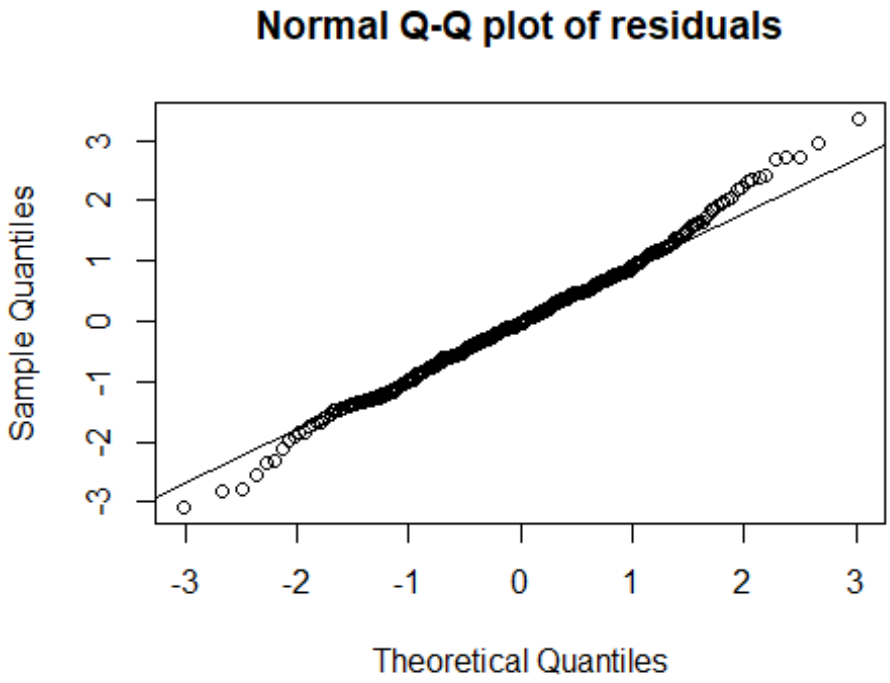
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1613    2225    2804    2978    3615    5140
```



Answer: See the plot above.

(7) Construct a scatterplot of semi-studentized residuals against the predicted values and a normal probability plot of semi-studentized residuals. Perform model diagnostics using the two plots.





Answer:

A scatterplot of semi-studentized residuals against the predicted value can show three things: (1) linearity (linear relationship between X and Y'), (2) constant variance of error terms, and (3) presence of outliers. The scatterplot shows fairly random plot, which suggests no violation on linearity. In addition, if you draw two horizontal lines on the scatterplot, the width from left to right is not so different, indicating no violation of constant variance of error terms assumption. Lastly, there are no observations in which |e*| > 4 so there seems to be no outliers. Thus, based on the scatterplot, there is no suggestion of violations of assumptions.

The normal probability plot shows some points on the far left side and far right side departing from the linear line, but the departure is not significant and most of the scatter points fall on the straight line. This suggests the normal assumption for the error terms is not violated .Therefore, the normal probability plot below does not indicate any violation of the normality assumption.

PART 2: MULTIPLE LINEAR REGRESSION ANALYSIS

This part aims at building a linear model using a set of potential predictor variables. In your report, make sure you include analyses for the following models.

Model #1: a regression model including three numerical predictor variables that you think are associated with the response variable.

Based on the result of the scatterplots and the summary results, we selected three numerical predictor variables of the following: V3: displacement, V4: horsepower, and V5: weight. In this part, we did not use any transformations on the original variables.

Since V5: weight was our independent variable in Part 1, we assigned V5(Weight) as X1 and V3(Displacement) and V4(Horsepower), the newly added variable, as X2, and X3 respectively. The response variable, Y, is V1(autoMPG).

In other words, Y = V1(AutoMPG), X1 = V5(Weight), X2 = V3(Displacement), and X3 = V4(Horsepower).

(i) Produce an ANOVA table. Report SST, SSR, and SSE, and their corresponding degrees of freedom.

```
##
## Call:
## lm(formula = V1 ~ V5 + V3 + V4, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3347  -2.8028  -0.3402   2.2037  16.2409
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.8559357  1.1959200  37.507  < 2e-16 ***
```

```
## V5          -0.0053516  0.0007124  -7.513 4.04e-13 ***
## V3          -0.0057688  0.0065819  -0.876  0.38132
## V4          -0.0416741  0.0128139  -3.252  0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.241 on 388 degrees of freedom
## Multiple R-squared:  0.707,  Adjusted R-squared:  0.7047
## F-statistic:   312 on 3 and 388 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Response: V1
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## V5          1 16497.8 16497.8 917.0641 < 2.2e-16 ***
## V3          1   150.9   150.9   8.3895  0.003988 **
## V4          1   190.3   190.3  10.5773  0.001245 **
## Residuals 388  6980.0    18.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 16839

## [1] 23819

## [1] 391
```

Answer:

SST: SSR + SSE = 16839 + 6980.0 = 23819 df(total): dfr + dfe = 3 + 388 = 391

SSR: 16497.8 + 150.9 + 190.3 = 16839 df(R): 1 + 1 + 1 = 3

SSE: 6980.0 df(E): 388

(ii) Perform the FF test of overall linear relationship. State the alternatives, the value of test statistic, p-value, and your conclusion.

```
## Analysis of Variance Table
##
## Response: V1
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## V5          1 16497.8 16497.8 917.0641 < 2.2e-16 ***
## V3          1   150.9   150.9   8.3895  0.003988 **
## V4          1   190.3   190.3  10.5773  0.001245 **
## Residuals 388  6980.0    18.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = V1 ~ V5 + V3 + V4, data = autompg1)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3347  -2.8028  -0.3402   2.2037  16.2409
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.8559357  1.1959200  37.507  < 2e-16 ***
## V5          -0.0053516  0.0007124  -7.513 4.04e-13 ***
## V3          -0.0057688  0.0065819  -0.876  0.38132
## V4          -0.0416741  0.0128139  -3.252  0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.241 on 388 degrees of freedom
## Multiple R-squared:  0.707,  Adjusted R-squared:  0.7047
## F-statistic:   312 on 3 and 388 DF,  p-value: < 2.2e-16
```

Answer:

(1) Alternatives: H_0: β 1 = β 2 = β 3 = 0 H_a: At least one β k ≠ 0

(2) value of test statistic, p-value, and conclusion:

The test statistic, F* is 312.

p-value is 2.2e-16.

Since the p-value is smaller than the alpha = 0.05 so we conclude Ha. In other words, we can conclude that at least one of the predictor variables have linear relationship with the response variable.

(iii) Compute the extra sum of squares SSR(X3|X1, X2) and the coefficient of partial determination R^2Y3|12.

```
## Analysis of Variance Table
##
## Model 1: V1 ~ V5 + V3
## Model 2: V1 ~ V5 + V3 + V4
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    389 7170.3
## 2    388 6980.0  1    190.28 10.577 0.001245 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 190.3

## Warning: package 'rsq' was built under R version 4.1.2

## $adjustment
## [1] FALSE
##
## $variables.full
```

```
## [1] "V5" "V3" "V4"
##
## $variables.reduced
## [1] "V5" "V3"
##
## $partial.rsq
## [1] 0.02653755
```

Answer:

extra sum of squares SSR(X3|X1, X2) : SSR(X3|X1, X2) = SSE(X1,X2) - SSE(X1,X2,X3) = 7170.3 - 6980.0 = 190.3

coefficient of partial determination R^2Y3|12: 0.02653755

(iv) Test whether X3 is helpful, given that X1 and X2 are in a model. State the alternatives, the value of test statistic, p-value, and your conclusion.

```
## Analysis of Variance Table
##
## Model 1: V1 ~ V5 + V3
## Model 2: V1 ~ V5 + V3 + V4
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    389 7170.3
## 2    388 6980.0  1    190.28 10.577 0.001245 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 10.57828

## [1] 3.865537
```

Answer:

(1) Alternatives: H_0: $\beta_3 = 0$ vs. H_a: $\beta_3 \neq 0$

(2) the value of test statistic, p-value, and conclusion: Test statisctic = F* = ((ssE(R)-SSE(F))/(dfR-dfF))/(SSE(F)/dfF) = ((7170.3-6980.0)/(389-388))/(6980.0/388) = 10.57828 p-value = 0.001245

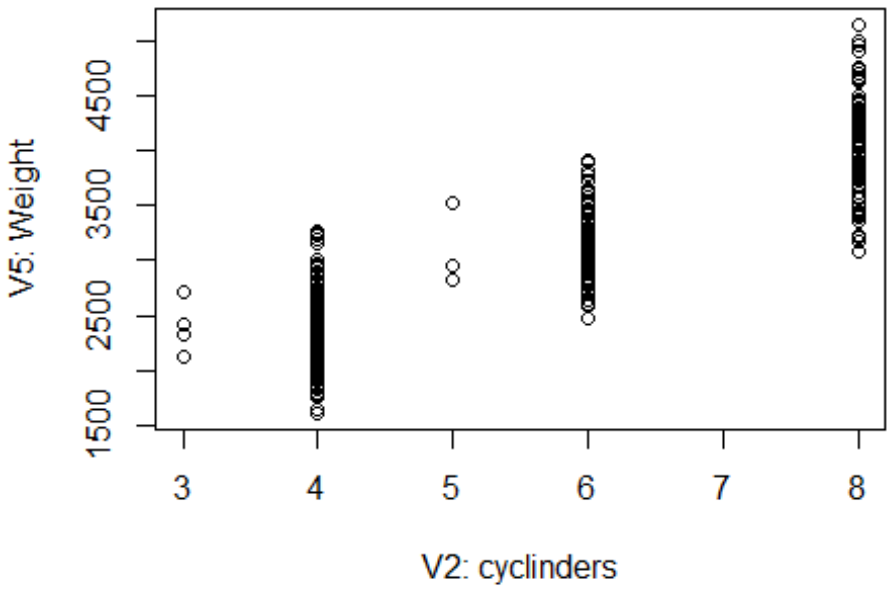
Conclusion: Since p-value is smaller than alpha = 0.05, we conclude Ha. In other words, we conclude the full model, which includes X3 variable (V4: Horsepower). Thus the test result shows that X3 is helpful given that X1 and X2 are in a model.

Model #2: a regression model including a numerical predictor variable, a categorical predictor variable, and their interaction term.

(v) Incorporate the categorical variable into the model by defining indicator variable(s).

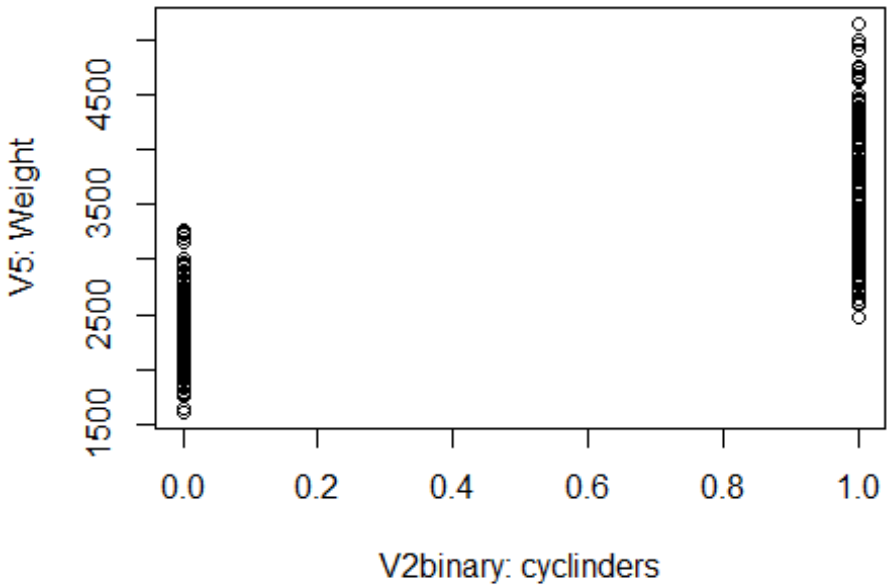
```
## [1] 8 4 6 3 5

##
## Call:
```

```
## lm(formula = V1 ~ V5 + V2 + V2 * V5, data = autompg1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4916  -2.6225  -0.3927   1.7794  16.7087
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.3864559  3.7333137  17.514  < 2e-16 ***
## V5          -0.0128348  0.0013628  -9.418  < 2e-16 ***
## V2          -4.2097950  0.7238315  -5.816 1.26e-08 ***
## V5:V2        0.0010979  0.0002101   5.226 2.83e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.165 on 388 degrees of freedom
## Multiple R-squared:  0.7174, Adjusted R-squared:  0.7152
## F-statistic: 328.3 on 3 and 388 DF,  p-value: < 2.2e-16

## [1] 0.8975273
```

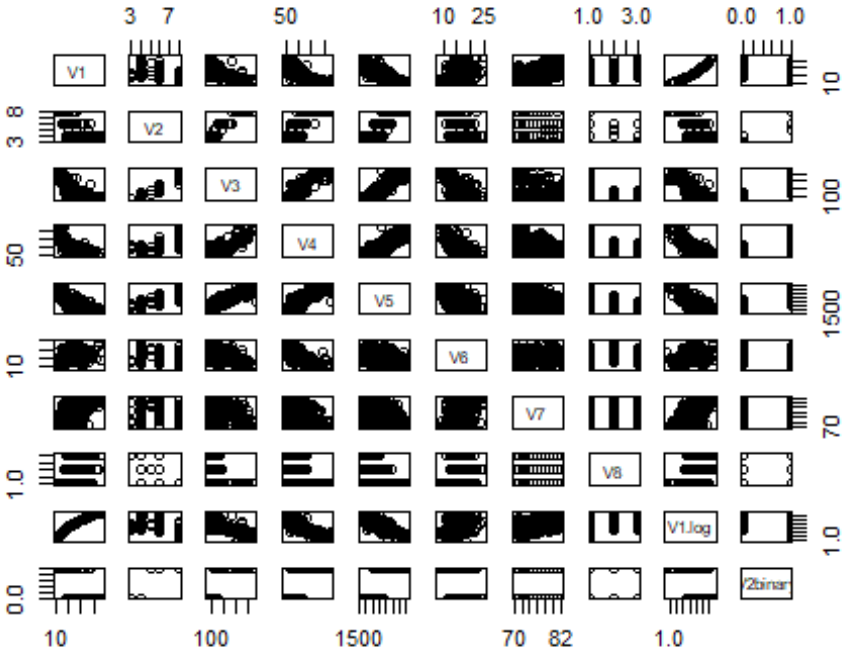



```
## Warning: package 'dplyr' was built under R version 4.1.1

##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## [1] 1 0

## [1] 0.8193009
```




```
##  [1] "V1"       "V2"       "V3"       "V4"       "V5"       "V6"
##  [7] "V7"       "V8"       "V5.exp"   "V5.inv"   "V1.log"   "V1.sqrt"
## [13] "V1.inv"   "V2binary"

##  [1] "V1"       "V2"       "V3"       "V4"       "V5"       "V6"
##  [7] "V7"       "V8"       "V1.log"   "V2binary"
```




```
##              V1    V2    V3    V4    V5    V6    V7    V8 V1.log V2binary
## V1         1.00 -0.78 -0.81 -0.78 -0.83  0.42  0.58  0.57   0.98    -0.75
## V2        -0.78  1.00  0.95  0.84  0.90 -0.50 -0.35 -0.57  -0.83     0.91
## V3        -0.81  0.95  1.00  0.90  0.93 -0.54 -0.37 -0.61  -0.85     0.85
## V4        -0.78  0.84  0.90  1.00  0.86 -0.69 -0.42 -0.46  -0.83     0.69
## V5        -0.83  0.90  0.93  0.86  1.00 -0.42 -0.31 -0.59  -0.88     0.82
## V6         0.42 -0.50 -0.54 -0.69 -0.42  1.00  0.29  0.21   0.45    -0.37
## V7         0.58 -0.35 -0.37 -0.42 -0.31  0.29  1.00  0.18   0.58    -0.29
## V8         0.57 -0.57 -0.61 -0.46 -0.59  0.21  0.18  1.00   0.56    -0.57
## V1.log     0.98 -0.83 -0.85 -0.83 -0.88  0.45  0.58  0.56   1.00    -0.77
## V2binary -0.75  0.91  0.85  0.69  0.82 -0.37 -0.29 -0.57  -0.77     1.00

##
## Call:
## lm(formula = V1 ~ V5 + V2binary + V2binary * V5, data = autompg11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7334  -2.5472  -0.3717   1.7696  17.1955
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.956e+01  1.998e+00  24.799  < 2e-16 ***
## V5          -8.863e-03  8.570e-04 -10.342  < 2e-16 ***
## V2binary    -1.355e+01  2.737e+00  -4.950 1.11e-06 ***
## V5:V2binary  3.820e-03  9.918e-04   3.851 0.000137 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.154 on 388 degrees of freedom
## Multiple R-squared:  0.7189, Adjusted R-squared:  0.7167
## F-statistic: 330.8 on 3 and 388 DF,  p-value: < 2.2e-16

## [1] 36.01

## [1] -0.005043
```

Answer:

We included the V5: Weight as the numerical predictor variable, X1, and V2: cyclinders as the categorical predictor variable, X2, and also included their interaction term, X1X2. This was based on examining the scatterplots and R square values generated in Part 1.

The possible values of the categorical variable X2 (V2 variable, cyclinders) are 3, 4, 5, 6, and 8. We first included the original variable, V2, without any modifications. However, because V5 has 5 possible values it is quite difficult to generate meaningful interpretations out of the result (see the below submodels and interpretation for the regression coefficinets). Also, we examined the scatterplot of V2 on V5 as well as the correlation coefficient and we find quite high correlation between the two variables.

$\hat{Y} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_{12} * X_1 X_2$

1. Unmodified Model

According to the R result, the fitted regression model is:

$\hat{Y}$ = 65.3864559 - 4.2097950 * X_1 -0.0128348 * X_2 + 0.0010979 * $X_1 X_2$

The possible values of the categorical variable X2 (V2 variable, cyclinders) are 3, 4, 5, 6, and 8.

Thus, we can generate 5 submodels :

1) When X2 = 3: $\hat{Y} = (\beta_0 + 3\beta_2) + (\beta_1 + 3\beta_{12}) * X_1$
2) When X2 = 4: $\hat{Y} = (\beta_0 + 4\beta_2) + (\beta_1 + 4\beta_{12}) * X_1$
3) When X2 = 5: $\hat{Y} = (\beta_0 + 5\beta_2) + (\beta_1 + 5\beta_{12}) * X_1$
4) When X2 = 6: $\hat{Y} = (\beta_0 + 6\beta_2) + (\beta_1 + 6\beta_{12}) * X_1$
5) When X2 = 8: $\hat{Y} = (\beta_0 + 8\beta_2) + (\beta_1 + 8\beta_{12}) * X_1$

β_2 has no practical meaning here because there are no observations with cars with cyclinders = 0.

β_{12} also has no practical meaning here.

So, for ease of analysis and interpretation, we decide to make modifications to the categorical variable, V2, by assigning 0 in case X2(=V2) $\leq$ 4 and 1 in case X2(=V2) > 4. In other words, the modified X2 will have the following values:

- X_2 = 0 if V2 = 3, 4
- X_2 = 1 if V2 = 5, 6, 8

2. Modified Model

According to the R result, the modified fitted regression model is as follows.

$\hat{Y} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_{12} * X_1 X_2$

$\hat{Y}$ = 49.56 - 0.008863 * X_1 - 13.55 * X_2 + 0.003820 * $X_1 X_2$

The possible values of the categorical variable X2(modified) are 0 and 1.

And according to the result, β_0 = 49.56, β_1 = - 0.008863, β_2 = - 13.55, β_{12} = 0.003820.

Thus, we can generate 2 submodels :

1) When X2 = 0:

   $\hat{Y} = \beta_0 + \beta_1 * X_1$

   = 49.56 - 0.008863 * X_1

2) When X2 = 1:

   $\hat{Y} = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) * X_1$

   = (49.56 + - 13.55) + (- 0.008863 + 0.003820) * X_1

   = 36.01 - 0.005043 * X_1

(vi) Interpret the coefficient of the categorical variable and that of the interaction term, respectively.

Answer:

β_2 is the increment in the value of the intercept when for the group with X2=1 compared with the group with X2=0. In other words, β_2 =-13.55 means that holding X1 constant, auto MPG is 13.55 lower for group with X2 = 1 compared with group with X2 =0.

β_{12} is the incremental slope of the for the group with X2=1. In other words, β_{12} = 0.003820 means that for every one-unit increase in weight(X1), auto MPG increases by 0.003820 more for group with X2 = 1 (cyclinders =5,6,8) compared with the corresponding increase in auto MPG for croup with X2 = 0 (cyclinders = 3.4).

(vii) Should you drop the interaction term? Explain.

```
## Analysis of Variance Table
##
## Model 1: V1 ~ V2binary + V5
## Model 2: V1 ~ V2binary + V5 + V2binary * V5
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    389 6951.4
## 2    388 6695.5  1    255.97 14.834 0.0001373 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

Testing H_0: $\beta_{12} = 0$ (Reduced model) vs. H_a: $\beta_{12} \neq 0$ (Full model)

The p-value = 0.0001373 < alpha = 0.05 so we conclude Ha. In other words, the full model, which includes the interaction term, is statistically better than the reduced model. So we should not drop the interaction term.

Model #3: building the "best" regression model using all potential predictor variables (with no interaction terms).

(viii) Use the AIC criterion and "backward elimination" procedure to obtain the "best" model. Report the subset of predictor variables to be included in the model and the corresponding AIC value.

```
##  [1] "V1"       "V2"       "V3"       "V4"       "V5"       "V6"
##  [7] "V7"       "V8"       "V1.log"   "V2binary"

## [1] "V1"       "V3"       "V4"       "V5"       "V6"       "V7"       "V8"
## [8] "V2binary"

## [1] 392   8

## Warning: package 'ALSM' was built under R version 4.1.1

## Loading required package: leaps

## Warning: package 'leaps' was built under R version 4.1.1

## Loading required package: SuppDists

## Warning: package 'SuppDists' was built under R version 4.1.1

## Loading required package: car

## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.1
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## [1] "numeric"

## Start:  AIC=919.64
## V1 ~ V2binary + V3 + V4 + V5 + V6 + V7 + V8
##
##            Df Sum of Sq    RSS     AIC
## - V6        1      5.57 3935.9  918.20
## <none>                  3930.3  919.64
## - V4        1     66.32 3996.6  924.20
## - V3        1    221.46 4151.8  939.13
## - V8        1    260.66 4191.0  942.82
## - V2binary  1    347.68 4278.0  950.87
## - V5        1    874.20 4804.5  996.37
## - V7        1   2334.52 6264.8 1100.41
##
## Step:  AIC=918.2
## V1 ~ V2binary + V3 + V4 + V5 + V7 + V8
##
##            Df Sum of Sq    RSS     AIC
## <none>                  3935.9  918.20
## - V4        1    144.40 4080.3  930.32
## - V3        1    215.91 4151.8  937.13
## - V8        1    260.93 4196.8  941.36
## - V2binary  1    350.95 4286.8  949.68
## - V5        1   1045.91 4981.8 1008.58
## - V7        1   2331.08 6267.0 1098.54

##
## Call:
## lm(formula = V1 ~ V2binary + V3 + V4 + V5 + V7 + V8, data = autompg2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9798 -1.9405 -0.2926  1.7698 12.9194
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.719e+01  3.954e+00  -4.347 1.77e-05 ***
## V2binary    -3.805e+00  6.495e-01  -5.859 1.00e-08 ***
## V3           2.749e-02  5.981e-03   4.596 5.86e-06 ***
## V4          -4.071e-02  1.083e-02  -3.758 0.000198 ***
## V5          -5.641e-03  5.577e-04 -10.115  < 2e-16 ***
## V7           7.378e-01  4.886e-02  15.100  < 2e-16 ***
## V8           1.344e+00  2.660e-01   5.052 6.76e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.197 on 385 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.8322
## F-statistic: 324.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

Answer:

According to the R result using AIC criterion and "backward elimination" procedure, the best model should include all variables except V6: Acceleration. The fitted model is as follows. The corresponding AIC value is 918.2.

$\hat{Y}$ = -17.19 - 3.805 * X_1 + 0.02749 * X_2 - 0.04071 * X_3 + - 0.005641 * X_4 + 0.7378 * X_5 + 1.344 * X_6

where

Y = auto MPG

X_1 = 0 if cyclinders = 3, 4; 1 if cyclinders = 5, 6, 8

X_2 = Displacement

X_3 = Horsepower

X_4 = Weight

X_5 = Model Year

X_6 = Origin