Chapter 3 Applied questions

Dr. Lasanthi Watagoda

Last compiled: Mar 15, 2018

Q8. This question involves the use of simple linear regression on the `Auto` data set.

(a) Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output.

```
library(ISLR)
attach(Auto)
summary(Auto)
```

```
      mpg          cylinders      displacement     horsepower
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0

     weight      acceleration        year           origin
 Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
 1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
 Median :2804   Median :15.50   Median :76.00   Median :1.000
 Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
 3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
 Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000

                 name
 amc matador       :  5
 ford pinto        :  5
 toyota corolla    :  5
 amc gremlin       :  4
 amc hornet        :  4
 chevrolet chevette:  4
 (Other)           :365
```

```
model1 <- lm(mpg ~ horsepower)
summary(model1)
```

```
Call:
lm(formula = mpg ~ horsepower)

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 ***
horsepower  -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
```

i. Is there a relationship between the predictor and the response?

Need to check the following hypothesis to answer this question. Conduct the four step test.

1. $H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$
2. t-statistic = -24.49
3. p-value = <2e-16
4. Conclusion: P-value $< \alpha$, Reject H_0, There is a relationship between horsepower and mpg.

ii. How strong is the relationship between the predictor and the response?

$R^2 = 0.6059$, 60.59% of the variation in the mpg is explained by horsepower.

Percentage error $= \frac{\text{RSE}}{\text{Mean of the response}} \times 100\% = \frac{4.906}{23.45} \times 100\% = 20.9211\%$

iii. Is the relationship between the predictor and the response positive or negative?

The coeffient of `horsepower` is negative (-0.157845) Therefore the relationship between `mpg` and `horsepower` is negative.

The more `horsepower` an automobile has the less `mpg` fuel efficiency the automobile will have.

iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

```
#predict(model1, data.frame(horsepower=c(98)))
predict(model1, data.frame(horsepower=c(98)), interval="confidence")
```

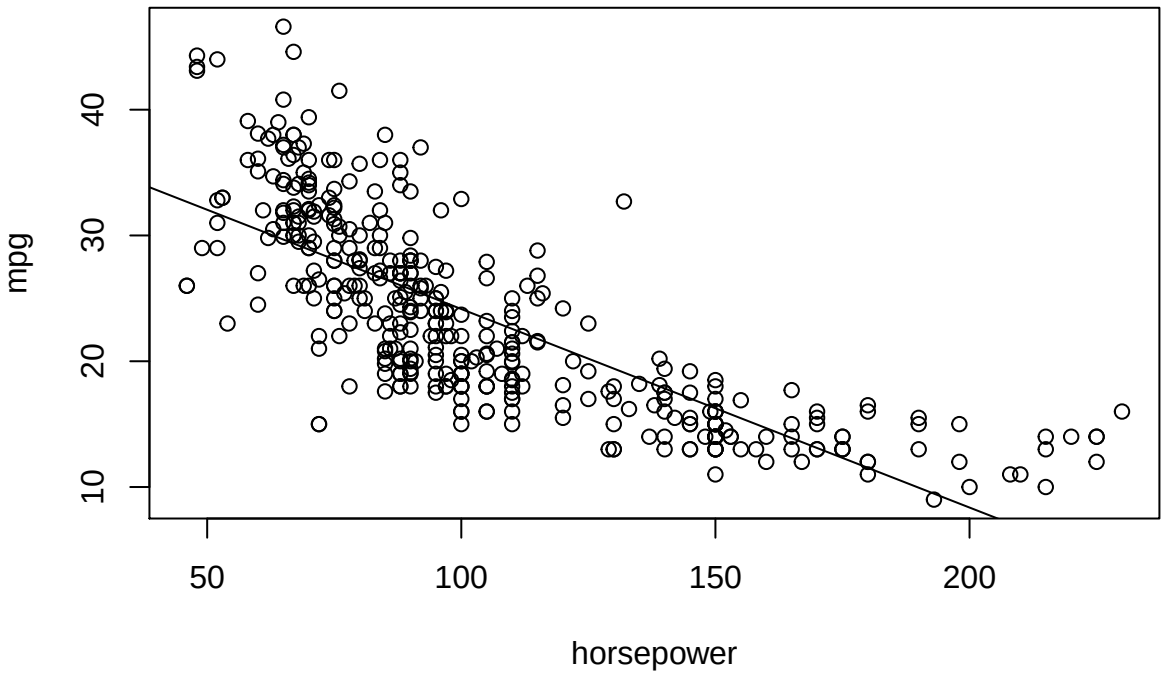
```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

```
predict(model1, data.frame(horsepower=c(98)), interval="prediction")
```

```
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

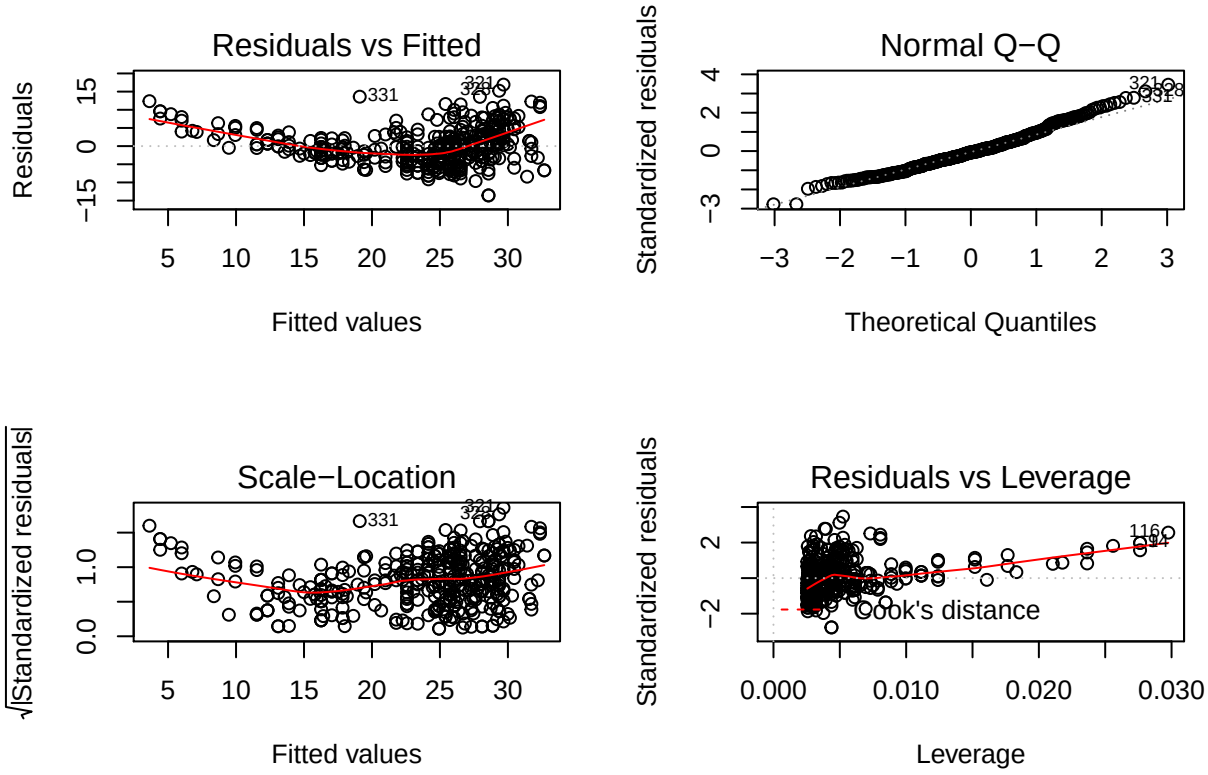
(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

```
plot(horsepower, mpg)
abline(model1) #add the line from the model to this plot
```



(c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```
par(mfrow=c(2,2))
plot(model1)
```

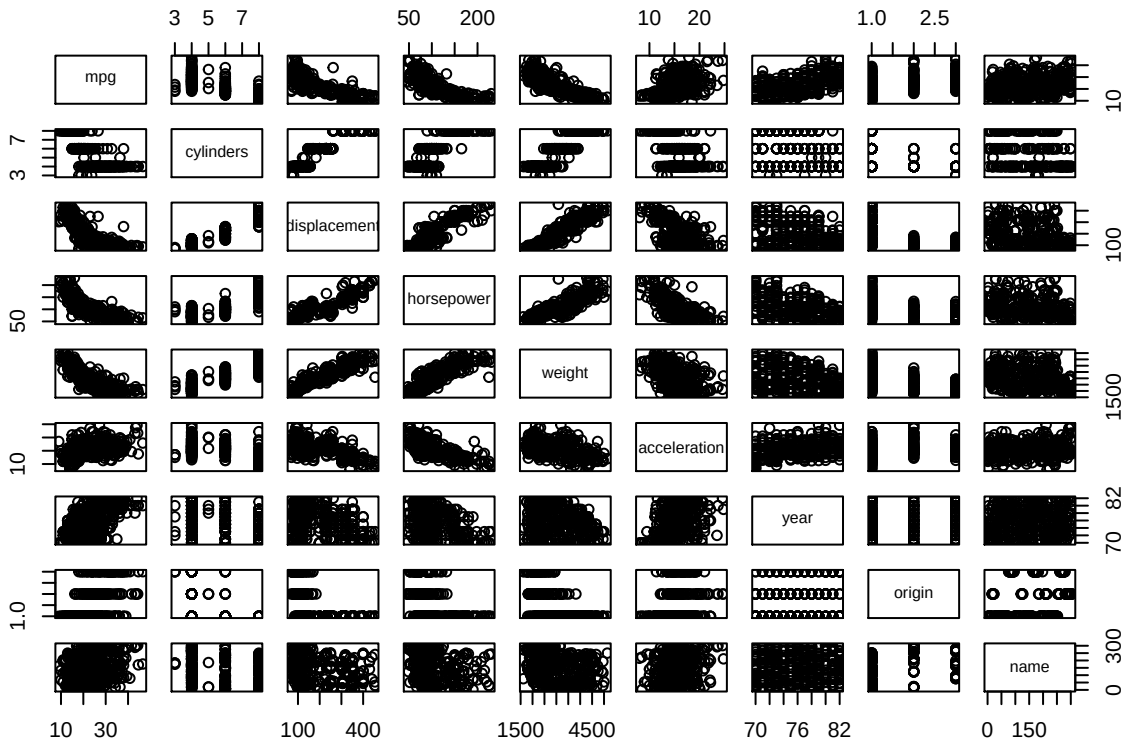


The relationship between mpg and horsepower is not linear. The residual plot recognizes this non linearity.

Q9. This question involves the use of multiple linear regression on the Auto data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
pairs(Auto)
```



(b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

```
head(Auto)
```

```
  mpg cylinders displacement horsepower weight acceleration year origin
1  18         8          307        130   3504         12.0   70      1
2  15         8          350        165   3693         11.5   70      1
3  18         8          318        150   3436         11.0   70      1
4  16         8          304        150   3433         12.0   70      1
5  17         8          302        140   3449         10.5   70      1
6  15         8          429        198   4341         10.0   70      1
                       name
1 chevrolet chevelle malibu
2         buick skylark 320
3        plymouth satellite
4             amc rebel sst
5               ford torino
6          ford galaxie 500
```

```
cor(Auto[,-c(9)])
```

```
                    mpg  cylinders displacement horsepower     weight
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
```

```
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
             acceleration       year     origin
mpg             0.4233285  0.5805410  0.5652088
cylinders      -0.5046834 -0.3456474 -0.5689316
displacement   -0.5438005 -0.3698552 -0.6145351
horsepower     -0.6891955 -0.4163615 -0.4551715
weight         -0.4168392 -0.3091199 -0.5850054
acceleration    1.0000000  0.2903161  0.2127458
year            0.2903161  1.0000000  0.1815277
origin          0.2127458  0.1815277  1.0000000
```

(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output.

```
model2 <- lm(mpg ~ .-name,data = Auto)
summary(model2)
```

```
Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

$$\hat{mpg} = -17.218435 - 0.493376 cylinders + 0.019896 displacement - 0.016951 horsepower - 0.006474 weight + 0.080576 acceleration + 0.750773 year + 1.426141 origin$$

i. Is there a relationship between the predictors and the response

Need to check the following hypothesis to answer this question. Conduct the four step test.

1. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0 \quad vs. \quad H_1$: At least one β_j is non zero 2.F-statistic = 252.4
2. p-value = <2.2e-16
3. Conclusion: P-value $< \alpha$, Reject H_0

ii. Which predictors appear to have a statistically significant relationship to the response?

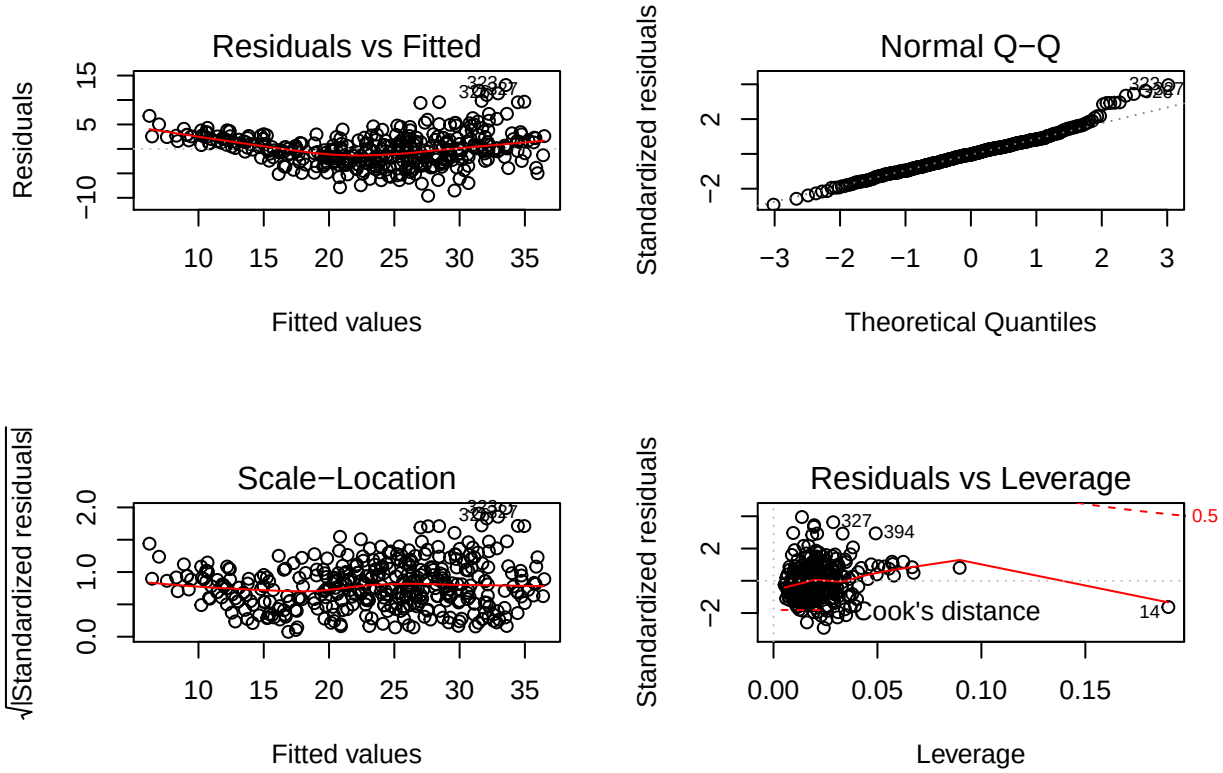
P-values for displacement, weight, year, and origin are smallar than 0.05 indicate that they are significant predictors that can be used to predict mpg.

iii. What does the coefficient for the year variable suggest?

For every one year, mpg increases by the coefficient(0.750773) holding all other predictors fixed.

(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit.

```
par(mfrow=c(2,2))
plot(model2)
```



1. **Residuals vs Fitted**

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

2. **Normal Q-Q**

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.

3. **Scale-Location**

It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.

4. **Residuals vs Leverage**

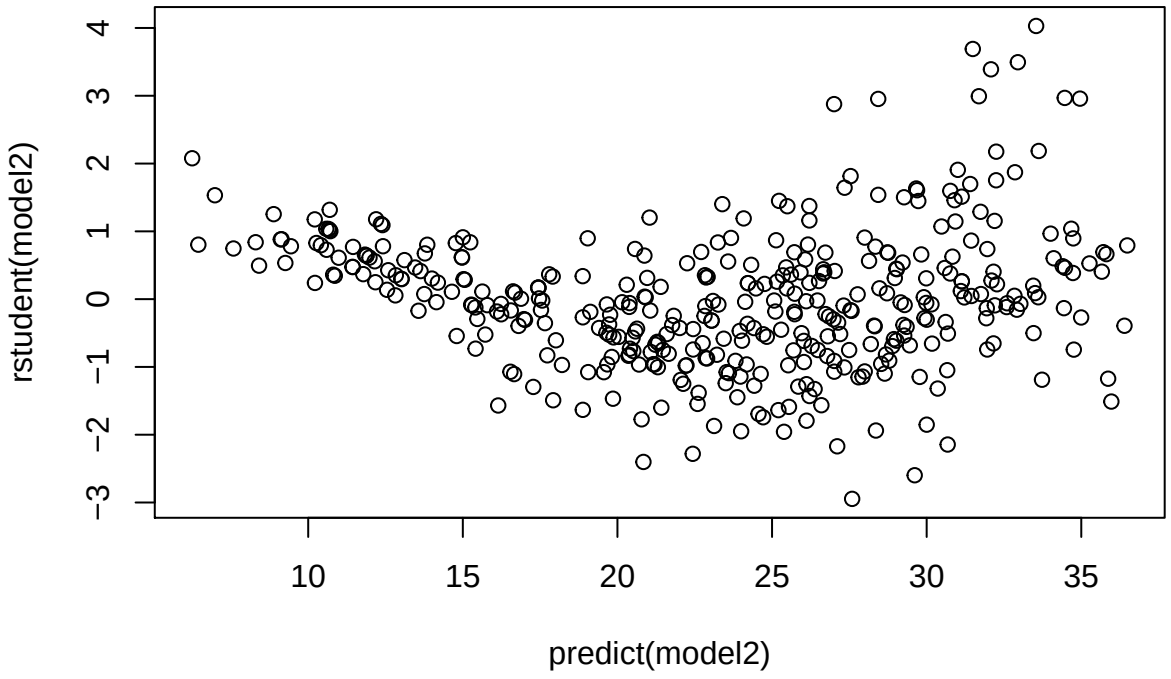
This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be

influential to determine a regression line. That means, the results wouldn't be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don't really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases.

Do the residual plots suggest any unusually large outliers? Yes.

Does the leverage plot identify any observations with unusually high leverage?

```
plot(predict(model2), rstudent(model2))
```



There are influential points (points with a value greater than 3)

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
model3 = lm(mpg~cylinders*displacement+displacement*weight)
summary(model3)
```

```
Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2934  -2.5184  -0.3476   1.8399  17.7723

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders            7.606e-01  7.669e-01   0.992    0.322
displacement        -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
```

```
weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,    Adjusted R-squared:  0.7237
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

Interaction between displacement and weight is statistically signifcant (P-value = 2.64e-05) Interactiion between cylinders and displacement is not statistically signifcant (P-value = 0.384).

(f) Try a few different transformations of the variables, such as $log(X)$, $\sqrt{X}$, X^2. Comment on your findings.

```
model4 = lm(mpg~sqrt(weight)+log(horsepower)+acceleration+I(acceleration^2))

summary(model4)
```

```
Call:
lm(formula = mpg ~ sqrt(weight) + log(horsepower) + acceleration +
    I(acceleration^2))

Residuals:
     Min       1Q   Median       3Q      Max
-10.7338  -2.4273  -0.1604   2.1804  15.5506

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       122.51698    9.29220  13.185  < 2e-16 ***
sqrt(weight)       -0.44050    0.06744  -6.531 2.05e-10 ***
log(horsepower)   -12.42799    1.92756  -6.448 3.39e-10 ***
acceleration       -1.97369    0.60125  -3.283  0.00112 **
I(acceleration^2)   0.04986    0.01788   2.788  0.00556 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.97 on 387 degrees of freedom
Multiple R-squared:  0.7439,    Adjusted R-squared:  0.7412
F-statistic:   281 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
?I
```