DSO 530 - Applied Modern Statistical Learning Methods

Homework 01

Hafsa Dawood, USC ID: 6829-4732-79

February 4, 2018

(1) Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

Table 3.4:

Variable	Coefficient	Std. error	t-stat	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.01	0.0059	-0.18	0.8599

The null hypotheses associated with Table 3.4 are that advertising budgets of "TV", "radio" or "newspaper" do not have an effect on sales correspondingly shown as: ${H_0}^{(1)} : \beta_1 = 0$; ${H_0}^{(2)} : \beta_2 = 0$; ${H_0}^{(3)} : \beta_3 = 0$. The corresponding p-values are highly significant for "TV" and "radio" and not significant for "newspaper". Hence we reject ${H_0}^{(1)}$ and ${H_0}^{(2)}$ and we do not reject ${H_0}^{(3)}$. So we can conclude that newspaper advertising budget do not affect sales.

(3) Suppose we have a data set with five predictors, $X_1 = GPA$, $X_2 = IQ$, $X_3 = Gender$ (1 for Female and 0 for Male), $X_4 = Interaction\ between\ GPA\ and\ IQ$, and $X_5 = Interaction\ between\ GPA\ and\ Gender$. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta_0} = 50$, $\hat{\beta_1} = 20$, $\hat{\beta_2} = 0.07$, $\hat{\beta_3} = 35$, $\hat{\beta_4} = 0.01$, $\hat{\beta_5} = -10$.

a. Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.
ii. For a fixed value of IQ and GPA, females earn more on average than males.
iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Solution:

Salary is given by

$$\hat{y} = 50 + 20(GPA) + 0.07(IQ) + 35(Gender) + 0.01(GPA\,X\,IQ) - 10(GPA\,X\,Gender)$$

For Males, $Gender = 0$ and for females, $Gender = 1$

Salary for Males

$$\hat{y} = 50 + 20(GPA) + 0.07(IQ) + 0.01(GPA\,X\,IQ) - 10(GPA\,X\,Gender)$$

Salary for Females

$$\hat{y} = 85 + 20(GPA) + 0.07(IQ) + 0.01(GPA\,X\,IQ) - 10(GPA\,X\,Gender)$$

So the starting salary for males is higher than for females. Hence (iii) is the right answer.

b. Predict the salary of a female with IQ of 110 and a GPA of 4.0. Solution:

$$\hat{y} = 85 + 20(4.0) + 0.07(110) + 0.01(4\,X\,110) - 10(4\,X\,1)$$

$$\hat{y} = 85 + 20(4.0) + 0.07(110) + 0.01(440) - 10(4)$$

```
y=85 + 20*4.0 + 0.07*110 + 0.01*440 - 10*4
y
```

```
## [1] 137.1
```

c. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

False. To verify if the GPA/IQ has an impact on the quality of the model we need to test the hypothesis $H_0 : \hat{\beta_4} = 0$ and look at the p-value associated with the t or the F statistic to draw a conclusion.

(4) I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

a. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer

Without having more details about the data, it is difficult to know if the RSS for linear regression is lower or the RSS for cubic regression. But it is assumed that the true relationship between X and Y is linear, we may expect the least squares line to be close to the true regression line, and hence the RSS for the linear regression may be lower than for the cubic regression.

b. Answer (a) using test rather than training RSS.

The test RSS depends upon the test data, so we have not enough information to conclude. However, we may assume that the cubic regression will have a higher test RSS as the overfit from training would have more error than the linear regression.

c. Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

If we assume that the true relationship between X and Y is not linear, then the cubic regression will have lower training RSS than the linear fit because of higher degree of flexibility which will reduce the training RSS.

d. Answer (c) using test rather than training RSS.

It is mentioned that it is not known how far from linear the relationship between X and Y is. If it is closer to linear than cubic, then linear regression test RSS could be lower. But if it is closer to cubic than linear, the cubic regression test RSS could be lower. This is due to bias and variance tradeoff: it is not clear what level of flexibility will fit the test data better, linear or cubic.

(5) Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y_i} = x_i \hat{\beta}$$

where

$$\hat{\beta} = (\sum_{i=1}^{n} x_i y_i)/(\sum_{i'=1}^{n} x_{i'}^2).$$

Show that we can write

$$\hat{y_i} = (\sum_{i'=1}^{n} a_i' y_i).$$

What is a_i' ?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

It is given that:

$$\hat{y_i} = x_i \hat{\beta}$$

where:

$$\hat{\beta} = (\sum_{j=1}^{n} x_j y_j)/(\sum_{k=1}^{n} {x_k}^2)$$

By replacing the value of $\hat{\beta}$ in the first equation we get:

$$\hat{y_i} = x_i \left(\frac{\sum_{j=1}^{n} x_j y_j}{\sum_{k=1}^{n} {x_k}^2} \right)$$

$$\hat{y_i} = \sum_{j=1}^{n} y_j \left(\frac{x_i x_j}{\sum_{k=1}^{n} {x_k}^2} \right)$$

Hence

$$\hat{y_i} = \sum_{j=1}^{n} a_j y_j$$

where

$$a_j = \left(\frac{x_i x_j}{\sum_{k=1}^{n} {x_k}^2} \right)$$

(6) Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

The least square line equation is:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Substituting x for $\bar{x}$

$$y = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$y = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}$$

$$y = \bar{y}$$

Hence the least square line passes through the point $(\bar{x}, \bar{y})$.

(7) It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

We know that:

$$R^2 = 1 - \left(\frac{RSS}{TSS}\right)$$

$$R^2 = 1 - \left(\frac{\sum_i (y_i - \hat{y_i})^2}{\sum_j {y_j}^2}\right)$$

Also we know:

$$\hat{y_i} = x_i \left(\frac{\sum_j x_j y_j}{\sum_j {x^2}_j}\right)$$

Replacing the value of $\hat{y_i}$ in the value of R^2

$$R^2 = 1 - \left(\frac{\sum_i \left(y_i - x_i \left(\frac{\sum_j x_j y_j}{\sum_j {x_j}^2}\right)\right)^2}{\sum_j {y_j}^2}\right)$$

$$R^2 = \left(\frac{\sum_j {y_j}^2 - \sum_{i=1} \left(y_i - x_i \left(\frac{\sum_j x_j y_j}{\sum_j {x_j}^2}\right)\right)^2}{\sum_j {y_j}^2}\right)$$

$$R^2 = \left(\frac{\sum_j {y_j}^2 - \sum_i \left({y_i}^2 - 2 y_i x_i \left(\frac{\sum_j x_j y_j}{\sum_j {x_j}^2}\right) + {x_i}^2 \left(\frac{\sum_j x_j y_j}{\sum_j {x_j}^2}\right)\right)}{\sum_j {y_j}^2}\right)$$

$$R^2 = \left(\frac{\sum_j {y_j}^2 - \sum_i {y_i}^2 + \sum_i 2 y_i x_i \left(\frac{\sum_j x_j y_j}{\sum_j {x_j}^2}\right) - \sum_i {x_i}^2 \left(\frac{\sum_j x_j y_j}{\sum_j {x_j}^2}\right)}{\sum_j {y_j}^2}\right)$$

$$R^2 = \left(\frac{2\left(\frac{(\sum_i x_i y_i)^2}{\sum_j {x^2}_j}\right) - \left(\frac{(\sum_i x_i y_i)^2}{\sum_j {x_j}^2}\right)}{\sum_j {y_j}^2}\right)$$

$$R^2 = \frac{\left(\frac{(\sum_i x_i y_i)^2}{\sum_j {x_j}^2}\right)}{\sum_j {y_j}^2}$$

$$R^2 = \frac{(\sum_i x_i y_i)^2}{\sum_j {x_j}^2 {y_j}^2}$$

$$R^2 = Cor(X, Y)^2$$

Hence proved.

(8) 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. For example :

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

The p-value corresponding to the F-statistic is very low < 0.05 indicating a relationship between "mpg" and "horsepower".

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.4.3
```

```
data = Auto
linear_reg1 <- lm(mpg ~ horsepower, data = data)
summary(linear_reg1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = data)
##
## 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
```

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

```
sigma(linear_reg1)/mean(data$mpg)
```

```
## [1] 0.2092371
```

To calculate the residual error relative to the response we use the mean of the response and the RSE. The mean of mpg is 23.44592. The RSE of the lm.linear_reg1 was 4.906 which indicates a percentage error of 20.92%. The R^2 is equal to 0.6059, almost 60.59% of the variability in "mpg" can be explained using "horsepower".

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

The coefficient of horsepower is -0.157, which is negative. So the relationship is also negative.

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

Predicted mpg for horsepower =98 and associated 95% confidence interval:

```
  predict(linear_reg1, data.frame(horsepower = 98), interval = "confidence")
```

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

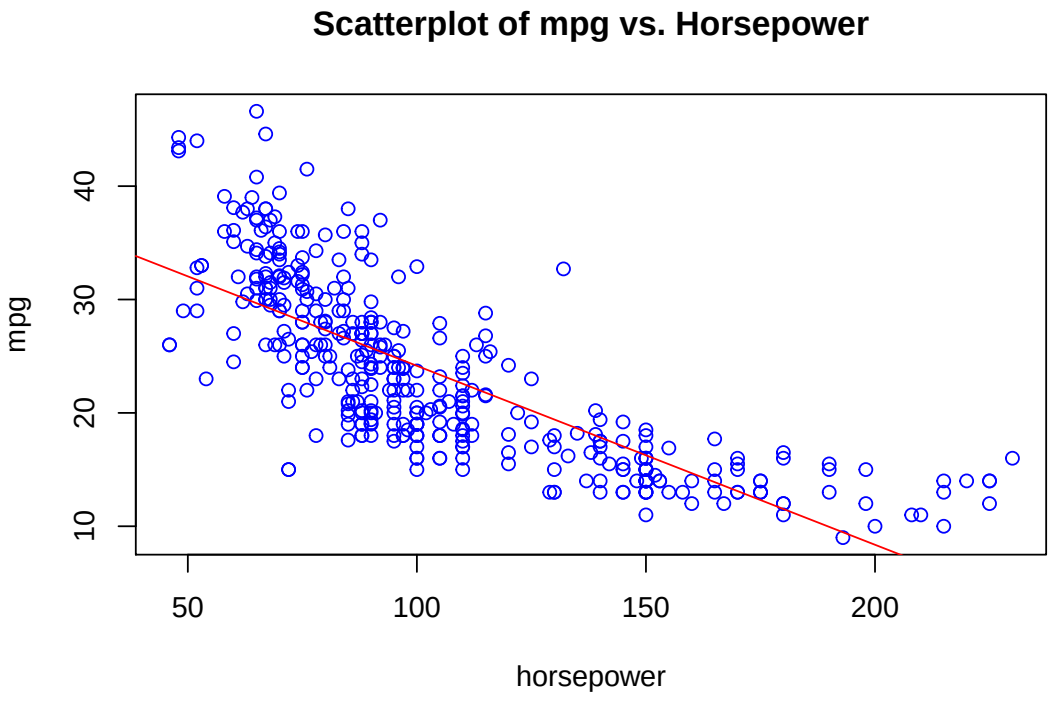
Predicted mpg for horsepower =98 and associated 95% prediction interval:

```
  predict(linear_reg1, data.frame(horsepower = 98), interval = "prediction")
```

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

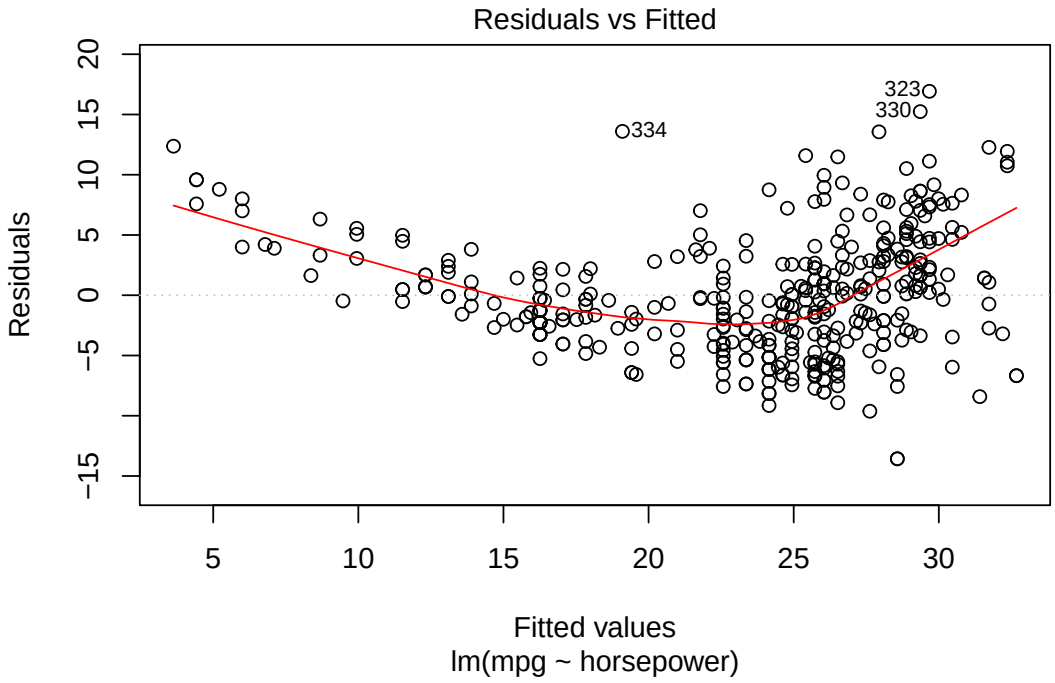
b. Plot the response and the predictor. Use the abline() function to display the least squares regression line.

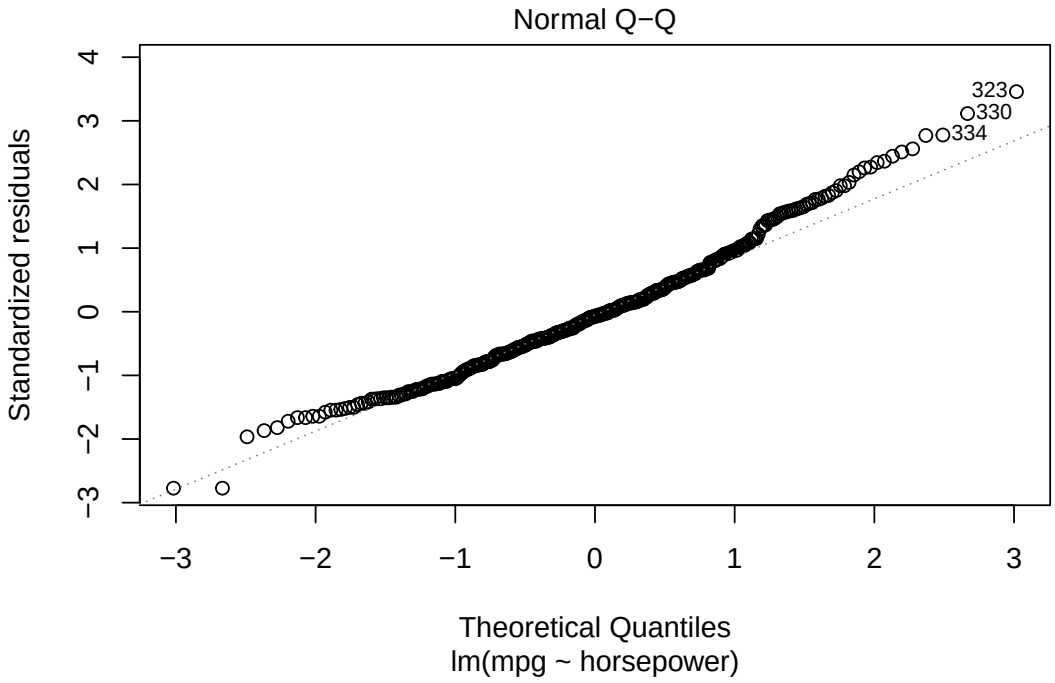
```
plot(data$horsepower, data$mpg, main = "Scatterplot of mpg vs. Horsepower",
       xlab = "horsepower", ylab = "mpg", col ="blue")
abline(linear_reg1, col = "red")
```

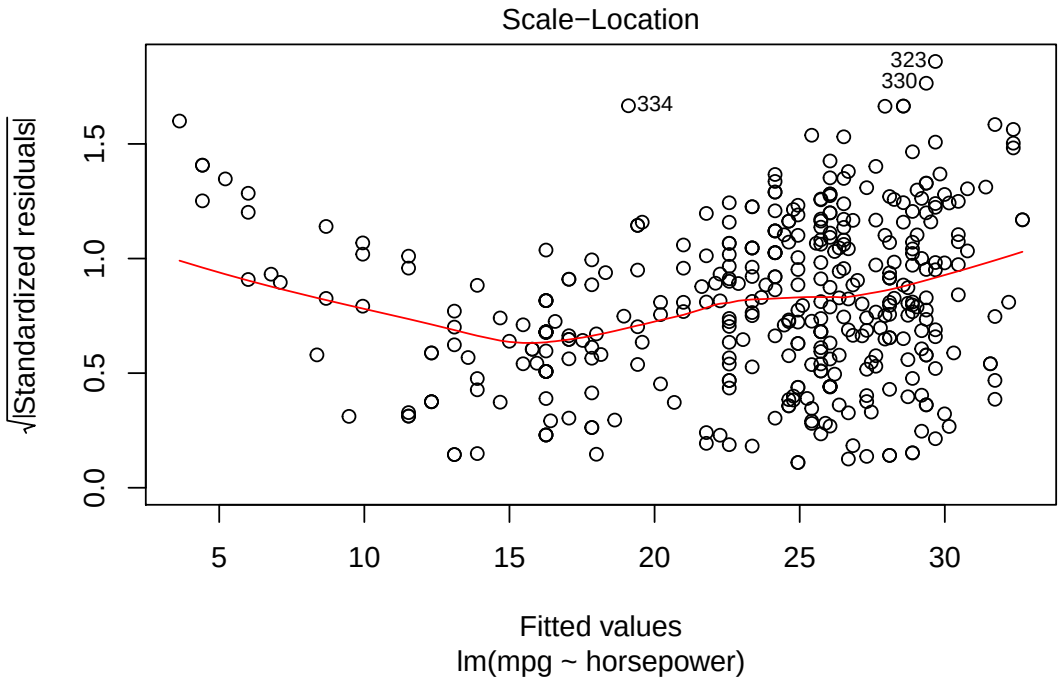


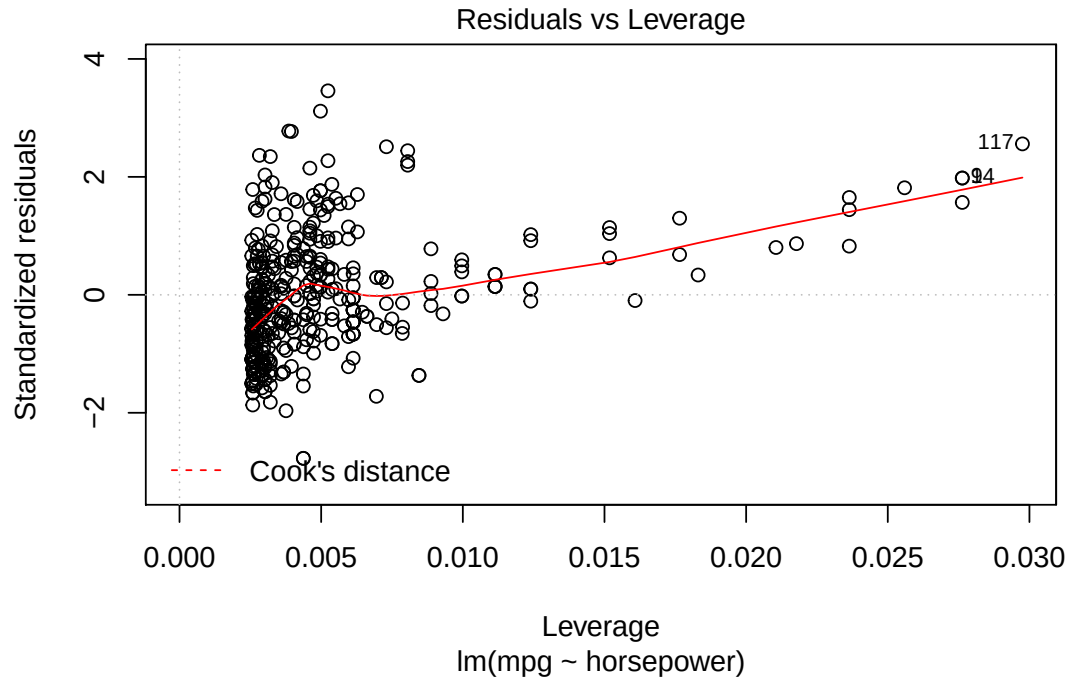
c. Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```
plot(linear_reg1)
```







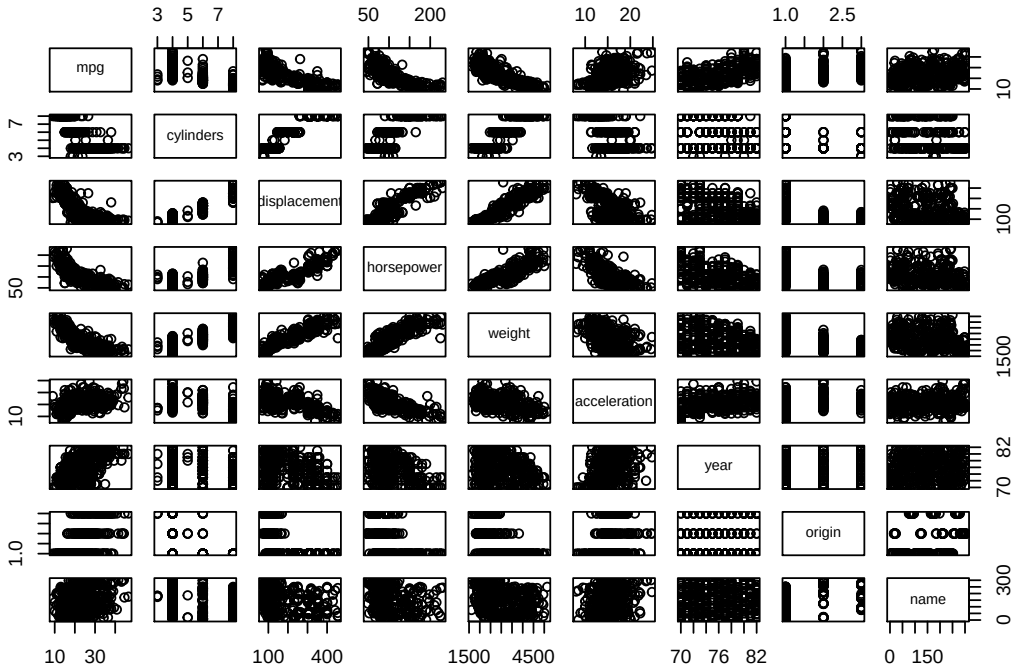


The plot of residuals versus fitted values indicates the presence of non linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (greater than 2 or lower than -2) and a few high leverage points.

(9) This question involves the use of multiple linear regression on the "Auto" data set.

a. Produce a scatterplot matrix which include all the variables in the data set.

```
library(ISLR)
data = Auto
pairs(data)
```



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

```
str(data)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54
```

```
cor(data[1:8])
```

```
##                     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. For instance :

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

To evaluathis we perform linear regression

```
linear_reg2 = lm(mpg ~ . - name, data = data)
summary(linear_reg2)
```

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

The p-value corresponding to the F-statistic is 2.037105910^{-139}, this indicates a clear evidence of a relationship between "mpg" and the other predictors.

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

On checking the p-values, all predictors are statistically significant except "cylinders", "horsepower" and "acceleration".

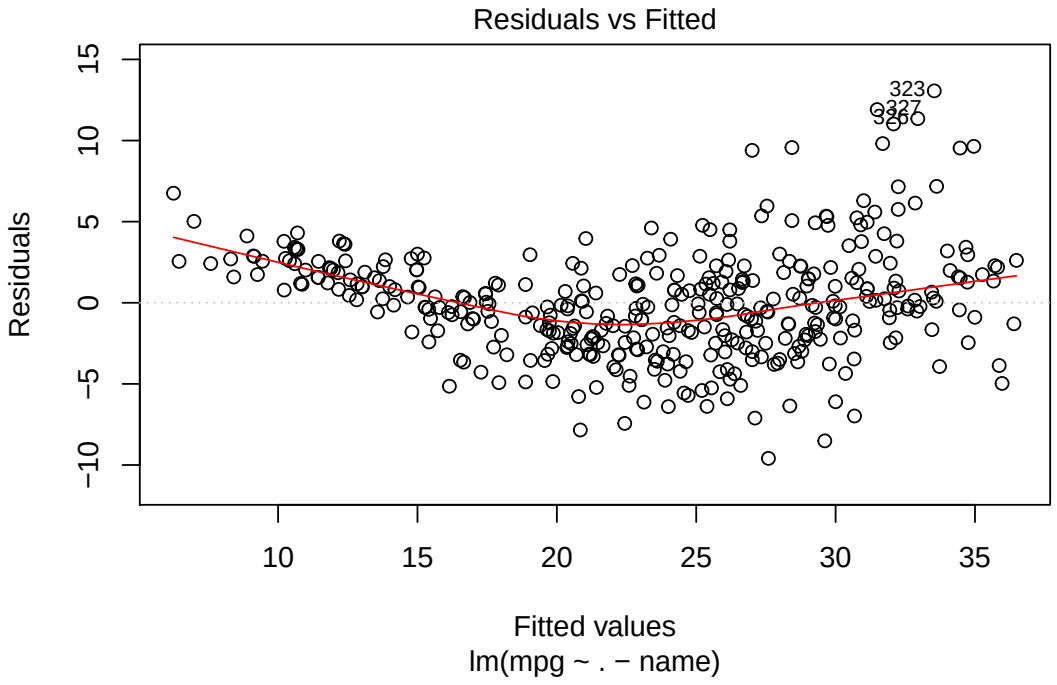
iii. What does the coefficient for the "year" variable suggest?

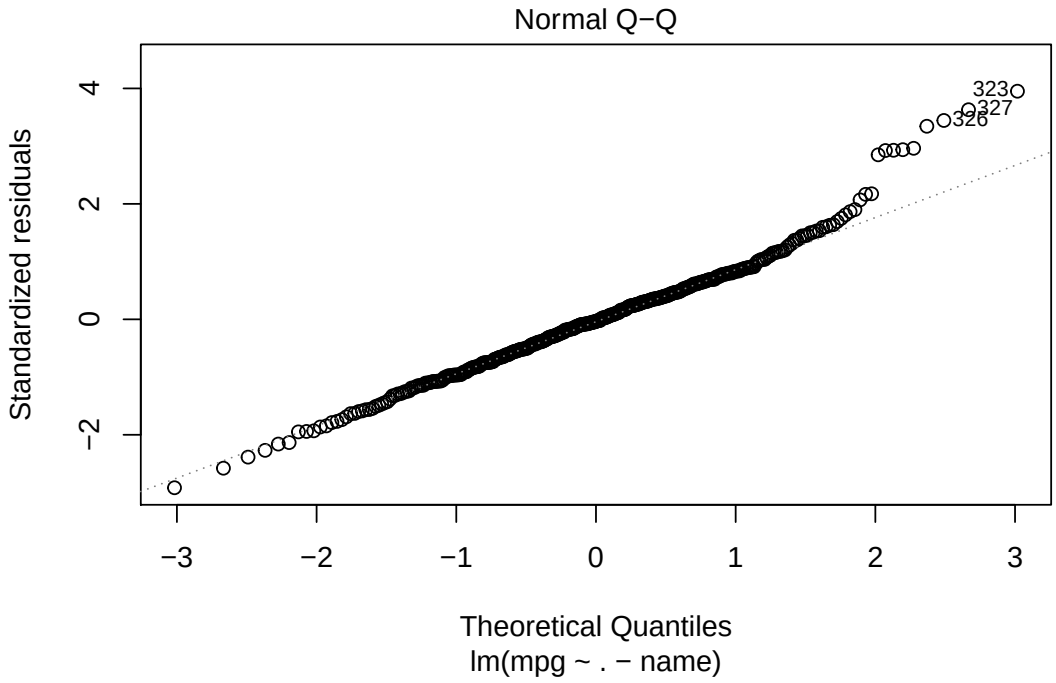
The coefficient ot the "year" variable suggests that the average effect of an increase of 1 year

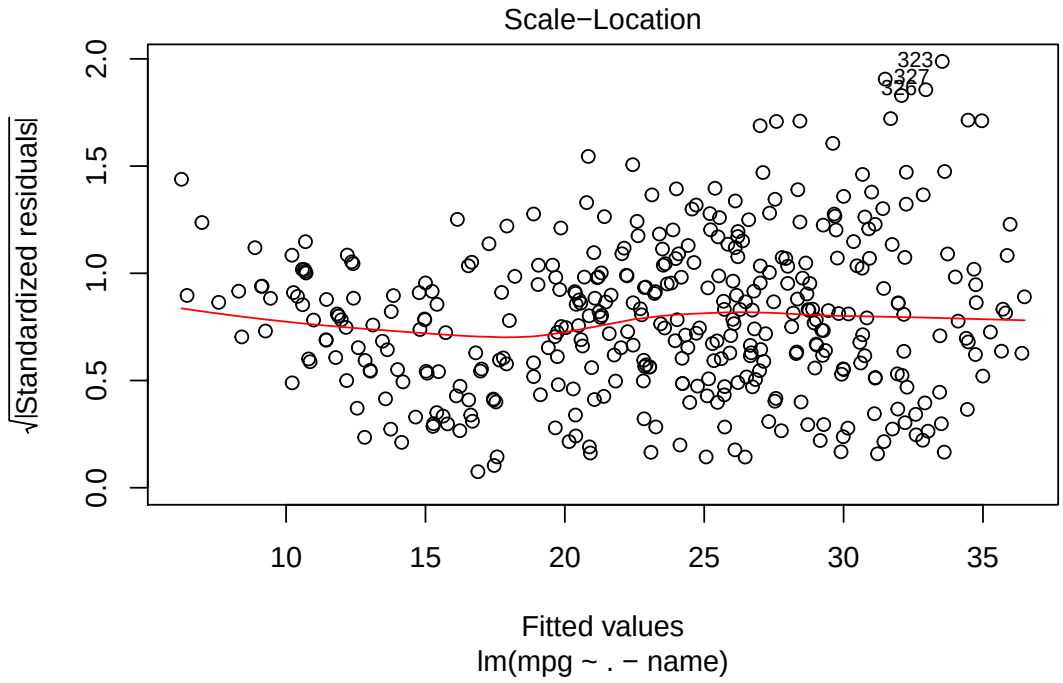
is an increase of 0.750773 in "mpg" (all other predictors remaining constant). In other words, cars become more fuel efficient every year by almost 1 mpg / year.

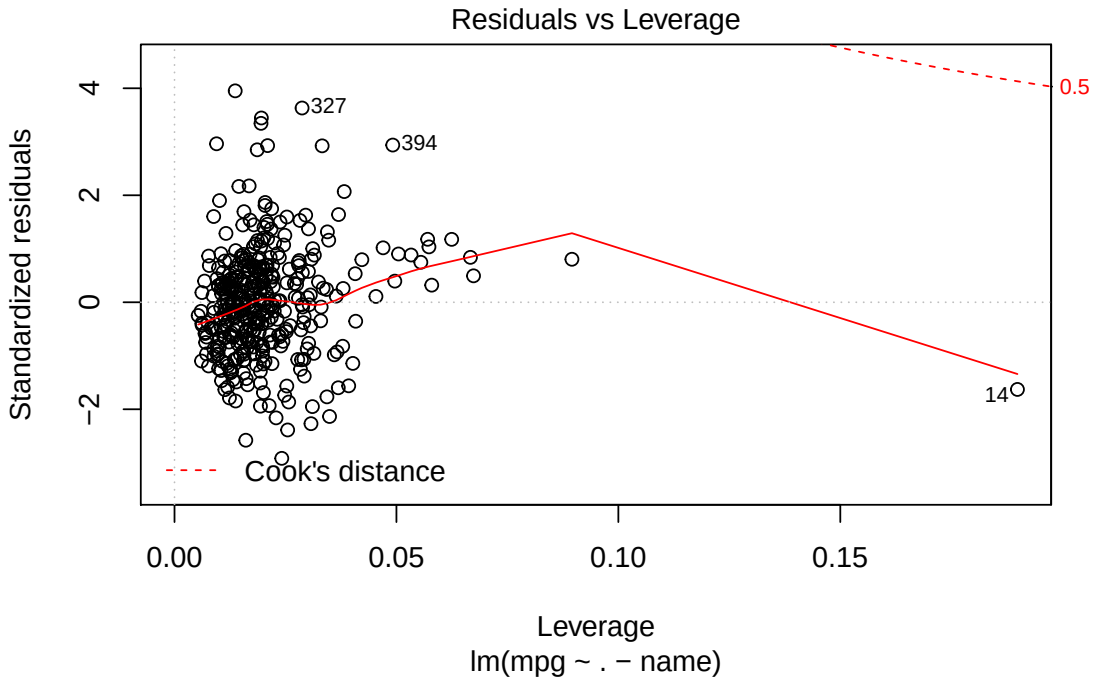
d. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers ? Does the leverage plots identify any observations with unusually high leverages?

```
plot(linear_reg2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ . − name)
323
327
326


Scale−Location
√|Standardized residuals|
Fitted values
lm(mpg ~ . − name)
323
327
326



The plot of residuals vs fitted values indicates slight non-linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (greater than 2 or less than -2) and one high leverage point (point 14).

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

When fitting with all predictors plus all possible interaction terms very few interactions appear statistically significant

```
linear_reg3.1 <- lm(mpg~.*.-name*.+.-name,data=data)
summary(linear_reg3.1)
```

```
##
## Call:
## lm(formula = mpg ~ . * . - name * . + . - name, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.548e+01  5.314e+01   0.668  0.50475
## cylinders                    6.989e+00  8.248e+00   0.847  0.39738
## displacement                -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                   5.034e-01  3.470e-01   1.451  0.14769
## weight                       4.133e-03  1.759e-02   0.235  0.81442
## acceleration                -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                         6.974e-01  6.097e-01   1.144  0.25340
## origin                      -2.090e+01  7.097e+00  -2.944  0.00345 **
```

```
## cylinders:displacement    -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower       1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight           3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration     2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year            -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin           4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower   -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight        2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year          5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin        2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight         -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year           -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin          2.233e-03  2.930e-02   0.076  0.93931
## weight:acceleration        2.346e-04  2.289e-04   1.025  0.30596
## weight:year               -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin             -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year          5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin        4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

From the correlation matrix, we obtained the two highest correlated pairs and used them in picking interaction effects. From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interactiion between cylinders and displacement is not.

```
linear_reg3.2 <- lm(mpg ~ cylinders * displacement+displacement * weight, data = data[, 1:8])
 summary(linear_reg3.2)
```

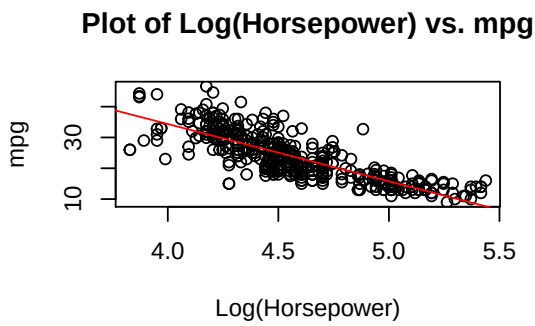
```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = data[, 1:8])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                           Estimate Std. Error t value 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
##
```

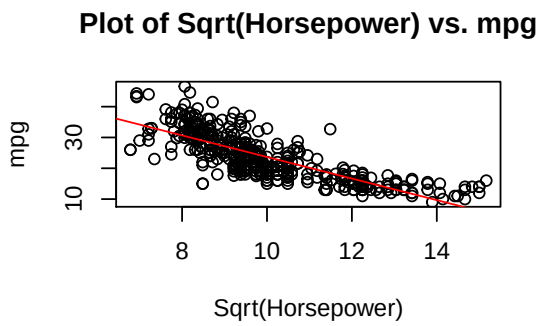
```
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

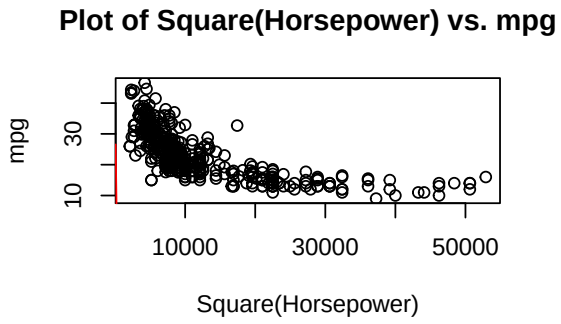
f. Try a few different transformations of the variables, such as $logX$, $\sqrt{X}$, X^2. Comment on your findings.

The log transformation gives the most linear plot.

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

plot(log(data$horsepower), data$mpg,
      main = "Plot of Log(Horsepower) vs. mpg", xlab = "Log(Horsepower)", ylab = "mpg")
abline(lm(mpg ~ log(horsepower), data = data), col = "red")

plot(sqrt(data$horsepower), data$mpg,
       main = "Plot of Sqrt(Horsepower) vs. mpg", xlab = "Sqrt(Horsepower)", ylab = "mpg")
abline(lm(mpg ~ sqrt(horsepower), data = data), col = "red")

plot((data$horsepower)^2, data$mpg,
      main = "Plot of Square(Horsepower) vs. mpg", xlab = "Square(Horsepower)", ylab = "mpg")
abline(lm(mpg ~ (horsepower)^2, data = data), col = "red")
```







(10) This question should be answered using the "Carseats" data set.

a. Fit a multiple regression model to predict "Sales" using "Price", "Urban" and "US".

```
library(ISLR)
data2 = Carseats
linear_reg4 = lm(Sales ~ Price + Urban + US, data = data2)
summary(linear_reg4)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b. Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative!

   i. The coefficient of the "Price" variable may be interpreted by saying that the average effect of a price increase of 1 dollar is a decrease of 54.4588492 units in sales all other predictors remaining fixed.
   ii. The coefficient of the "Urban" variable may be interpreted by saying that on average the unit sales in urban location are 21.9161508 units less than in rural location all other predictors remaining fixed.
   iii. The coefficient of the "US" variable may be interpreted by saying that on average the unit sales in a US store are 1200.5726978 units more than in a non-US store all other predictors remaining fixed.

c. Write out the model in equation form, being careful to handle the qualitative variables properly.

$$Sales = 13.043469 - 0.054459(Price) - 0.021916(Urban) + 1.200573(US) + \epsilon$$

If the store is in an urban location then Urban = 1 and if the store is in a non-urban location then Urban = 0. If the store is in the US then US = 1 and if the store is not in the US then US = 0

d. For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

We can reject the null hypothesis for the "Price" and "US" variables.

e. On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
linear_reg5 = lm(Sales ~ Price + US, data = data2)
summary(linear_reg5)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f. How well do the models in (a) and (e) fit the data?

Model (a) has an $AdjustedR^2 = 0.2335$ and $RSE = 2.472$. Model (e) has an $AdjustedR^2 = 0.2354$ and $RSE = 2.469$. Both models fit the data well but the model (e) has slightly lower RSE and better $AdjustedR^2$ than for the model (a).

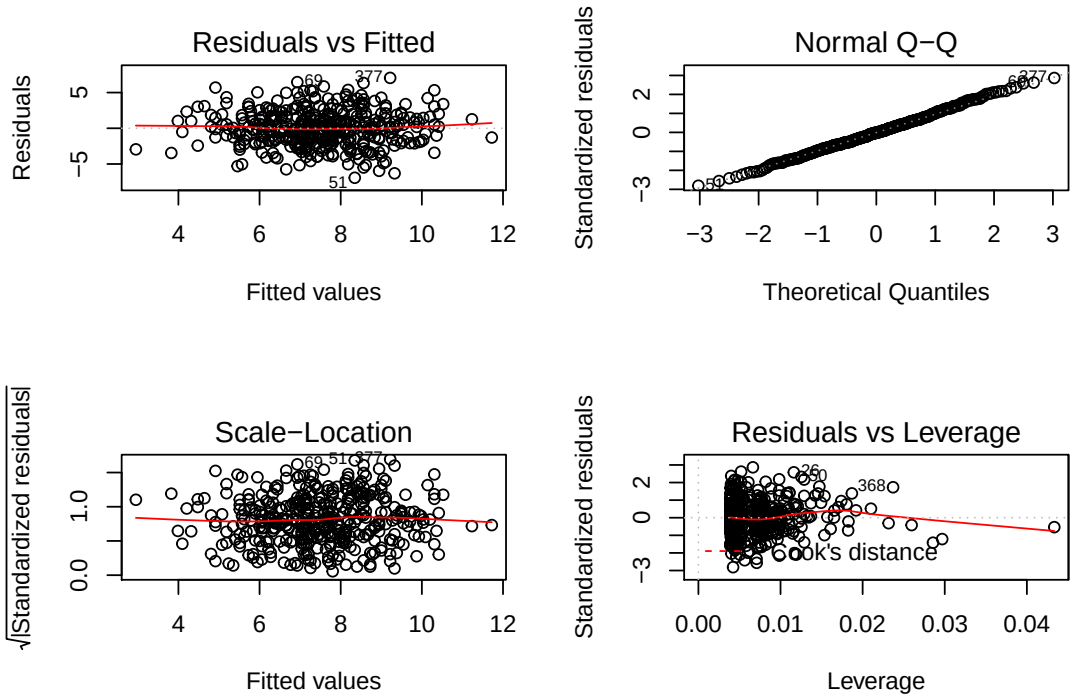
g. Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
confint(linear_reg5)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

h. Is there evidence of outliers or high leverage observations in the model from (e)?

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

plot(linear_reg5)
```



The plot of standardized residuals vs leverage indicates the presence of a few outliers (greater than 2 or lower than -2) and some leverage points.