Homework 1

Conceptual

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.

Answer:
Multiple regression coefficient estimates when TV, radio, and newspaper advertising budgets are used to predict product sales:

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon$$

three null hypothesis:
$H_0 : \beta_1 = 0$, in the presence of radio and newspaper ads, TV ads have no relationship with sales.
$H_0 : \beta_2 = 0$, in the presence of TV and newspaper ads, radio ads have no relationship with sales.
$H_0 : \beta_3 = 0$, in the presence of TV and radio ads, newspaper ads have no relationship with sales.

The low p-values(<0.0001) of TV and radio show that the null hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ are rejected for TV and radio, meaning TV and radio advertising budgets may make impacts on sales. While p-value of newspaper is high, suggesting that we do not reject null hypothesis $H_0 : \beta_3 = 0$ for newspaper. We may 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 forMale), $X_4 =$ 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 onaverage than males provided that the GPA is high enough.

Answer:
iii.For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
Regression equation:

$$\hat{y}_0 = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA * IQ - 10GPA * Gender$$

When Gender = 0 (Male):

$$\hat{y}_0 = 50 + 20GPA + 0.07IQ + 0.01GPA * IQ$$

When Gender = 1 (Female):

$$\hat{y}_0 = 85 + 10GPA + 0.07IQ + 0.01GPA * IQ$$

if:

$$50 + 20GPA85 + 10GPA$$

which is equivalent to $GPA 3.5$, males earn more on avearge than females. Therefore (iii.) is the right answer.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

Answer:
because $IQ = 110$, $Gender = 1$, and $GPA = 4$, $= 85 + 40 + 7.7 + 4.4 = 137.1$, $PredictedSalary = 137.1$

(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.

Answer:
False. To check if the GPA/IQ has a relationship with salary we need to test the hypothesis $H_0 : \hat{\beta}_4 = 0$ and look at the p-value to draw a conclusion. Because the p-value of the GPA/IQ interaction term is unknown, we cannot say the evidence of an interaction effect is little.

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.

Answer:
For the training data, it is difficult to tell. Cubic regression is a more flexible model, so the RSS in the training data is expected to be smaller compared with a linear regression. the RSS will decrease when adding more explanatory variables for training data, so the RSS for the cubic model will be lower than the RSS for the linear model.

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

Answer:
For test data, Linear regression correctly assumes the true data generating process, and the cubic model will be more overfitted than the linear model, so the RSS for the linear model will be lower than the RSS for the cubic model.

5

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

$$\hat{y}_i = x_i \hat{\beta}_i,$$

where

$$\hat{\beta} = (\sum_{i=1}^{n} x_i y_i)/(\sum_{i'=1}^{n} x_{i'}^2) \qquad (3.38)$$

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.*

Answer:

$$\hat{y}_i = x_i \hat{\beta}_i$$

$$y = x_i(\sum_{i=1}^{n} x_i y_i)/(\sum_{i'=1}^{n} x_{i'}^2) = x_i(\sum_{i'=1}^{n} x_i' y_i')/(\sum_{k=1}^{n} x_k^2) = \frac{\sum_{i'=1}^{n} x_i x_i' y_i'}{\sum_{k=1}^{n} x_k^2} = \sum_{i'=1}^{n} \frac{x_i x_i' y_i'}{\sum_{k=1}^{n} x_k^2} = \sum_{i'=1}^{n} \frac{x_i x_i'}{\sum_{k=1}^{n} x_k^2} y_i'$$

so

$$\hat{y}_i = \sum_{i'=1}^{n} a_i' y_i'$$

so

$$a_i' = \frac{x_i x_i'}{\sum_{k=1}^{n} x_k^2}$$

6

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

Answer:
From (3.4), we can know that least square line equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$
and $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
so $\hat{y} = \beta_0 + \beta_1 \bar{x} = \bar{y} - \beta_1 \bar{x} + -\beta_1 \bar{x} = \bar{y}$
the least squares line always 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 $\bar{x} = \bar{y} = 0$

Answer:
from (3.17), we know that: because $\bar{x} = \bar{y} = 0$

$$\begin{aligned} R^2 &= \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \frac{\sum_{i=1}^{n}(x_i - \bar{x_i})(y_i - \bar{y_i})}{\sum_{i=1}^{n}(x_i - \bar{x_i})^2} x_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sum_{i=1}^{n}(x_i)^2} x_i)^2}{\sum_{i=1}^{n}(y_i)^2} \end{aligned} \tag{1}$$

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(y_i)^2 - \sum_{i=1}^{n}(y_i - \frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sum_{i=1}^{n} x_i^2} x_i)^2}{\sum_{i=1}^{n}(y_i)^2} \\ &= \frac{2\sum_{i=1}^{n} x_i y_i \frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sum_{i=1}^{n} x_i^2} - \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{i=1}^{n} x_i^2}}{\sum_{i=1}^{n}(y_i)^2} \\ &= \frac{2 * \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{i=1}^{n} x_i^2} - \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{i=1}^{n} x_i^2}}{\sum_{i=1}^{n}(y_i)^2} \\ &= \frac{(\sum_{i=1}^{n} x_i y_i)^2 / \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(y_i)^2} \end{aligned} \tag{2}$$

so

$$R^2 = \frac{(\sum_{i=1}^n x_i y_i)^2}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2} = Cor(X,Y)^2$$

Proven.

Applied

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?
ii. How strong is the relationship between the predictor and the response?
iii. Is the relationship between the predictor and the response? positive or negative?
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

Answer:

```
library(ISLR)
data(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
```

```
library(ISLR)
lm.fit = lm(mpg ~ horsepower, data = Auto)
summary(lm.fit)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ horsepower, data = Auto)
##
## 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. Yes, there is a relationship between horsepower and mpg as deterined by testing the null hypothesis of all regression coefficients equal to zero. Since the p-value of the F-statistic is smaller than 0.05, we can reject the null hypothesis and conclude that there is a relationship.

ii. The adjusted R-squared of this model is 0.6049, which means 60.49% of the variance in the **mpg** can be explained by the variance in the predictor **horsepower**.

iii. The coeficient of "horsepower" is negative, so the relationship between the predictor and the response is negative.

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

```r
predict(lm.fit, data.frame(horsepower = 98), interval = 'confidence')
```

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

The predicted mpg is 24.47 associated with a horsepower of 98. Associated 95% confidence interval is [23.97, 24.96].

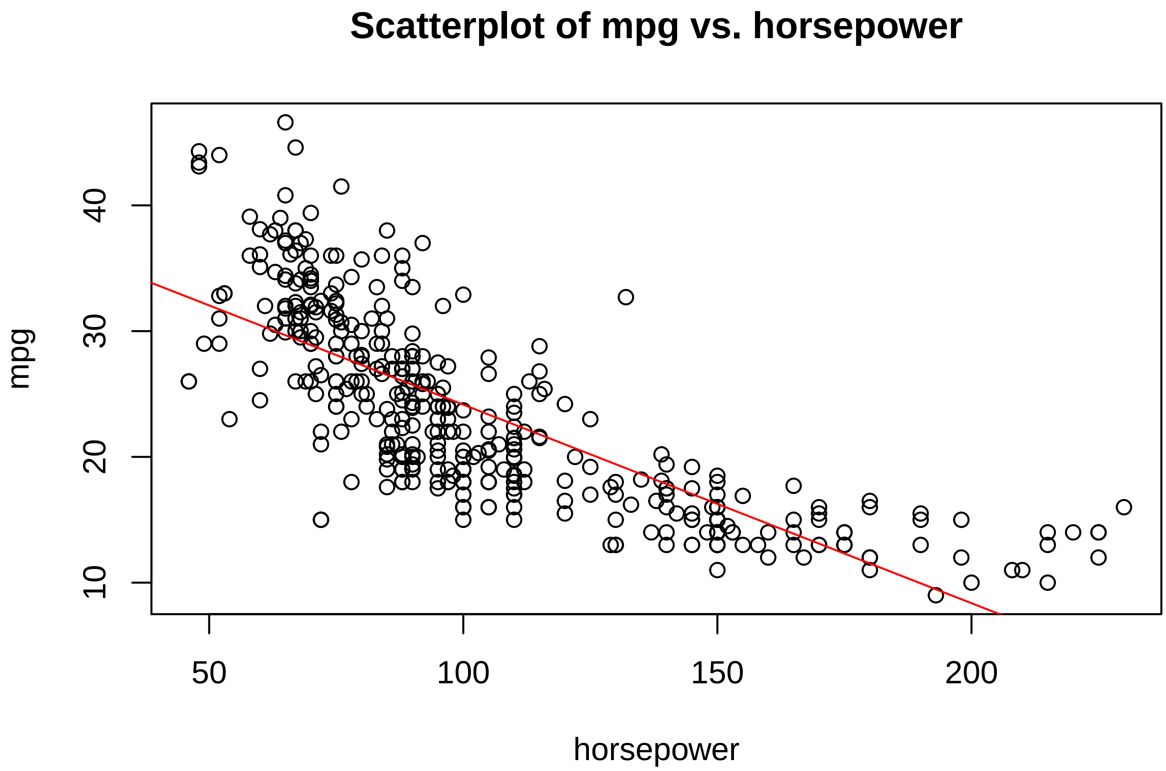
```r
predict(lm.fit, data.frame(horsepower = 98), interval = 'prediction')
```

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

Associated 95% prediction interval is [14.81, 34.12].

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

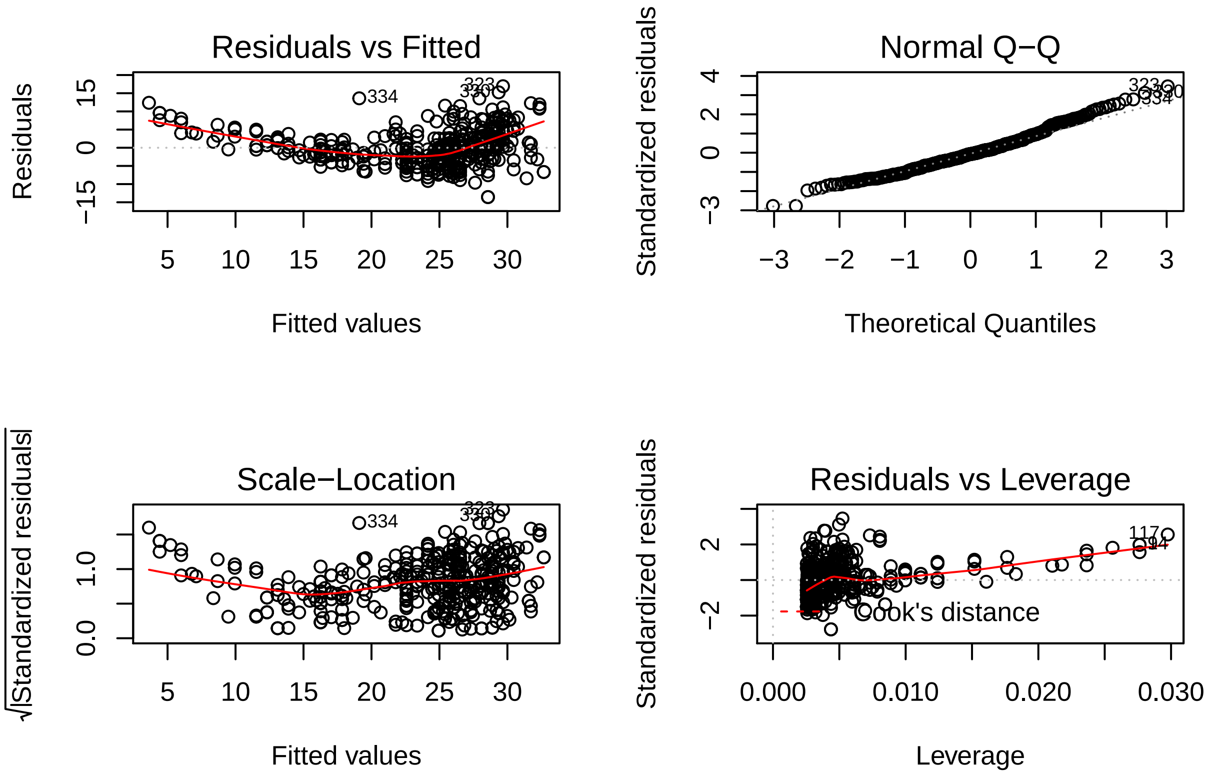
```r
plot(Auto$horsepower, Auto$mpg, main = "Scatterplot of mpg vs. horsepower",
     xlab = "horsepower", ylab = "mpg") +
abline(lm.fit, col = 'red')
```




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

(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(lm.fit)
```

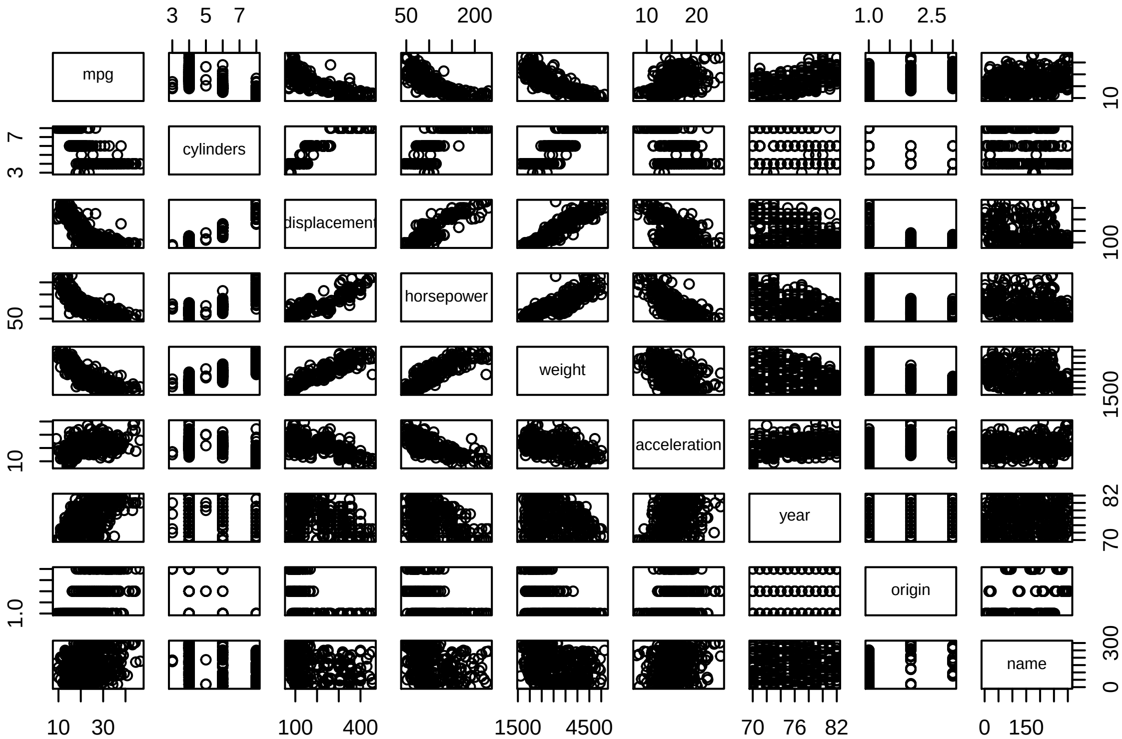


From the plot of residuals versus fitted values, we can find that there is some evidence of non-linearity. The plot of standardized residuals versus leverage shows there are some outliers.

9

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, cor() which is qualitative.

```
Auto_new = Auto[1:8]
cor(Auto_new)
```

```
##                     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?
ii. Which predictors appear to have a statistically significant relationship to the response?
iii. What does the coefficient for the year variable suggest?

Answer:

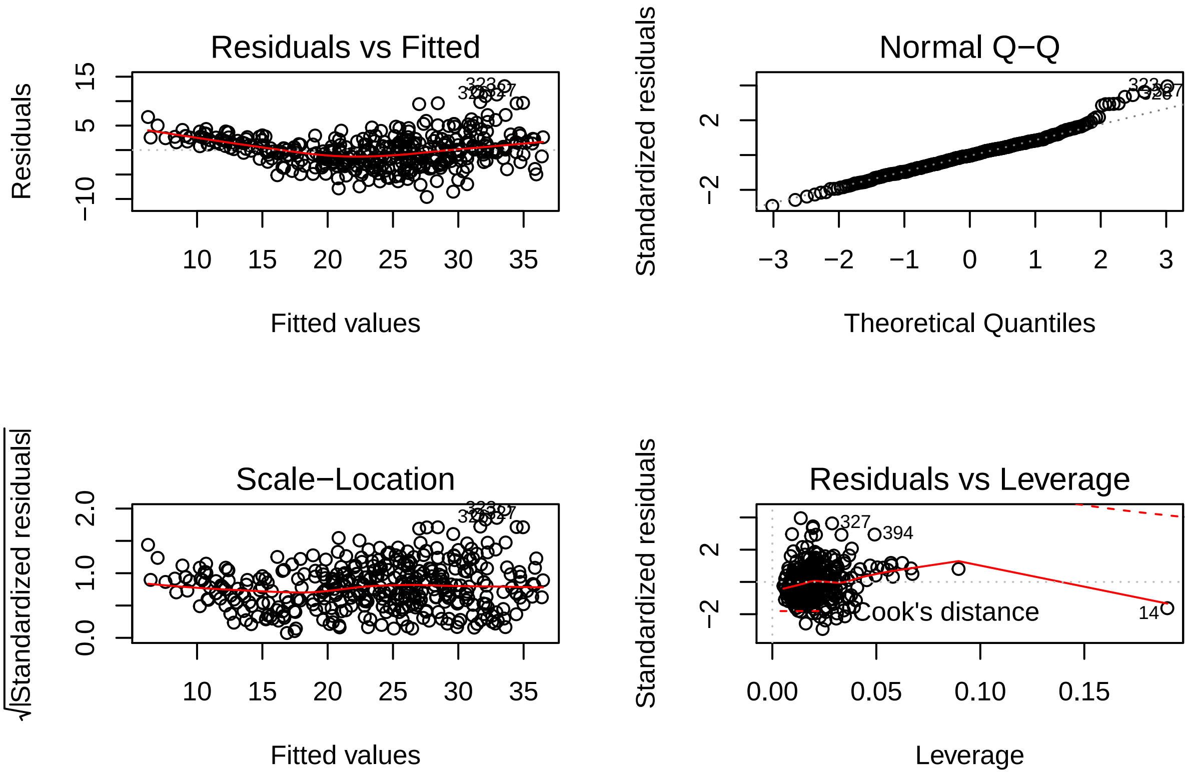
```r
lm.fit9 = lm(formula = mpg ~ ., data = Auto_new)
summary(lm.fit9)
```

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

i. Yes, there is a relationship between predictors and mpg. Since the p-value of the F-statistic is smaller than 0.05, we can reject the null hypothesis and conclude that there is a relationship.
ii. Displacement, weight, year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not.
iii. Given all other predictors remaining constant, the average effect of an increase of 1 year is an increase of 0.7507727 in "mpg".

(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 plot identify any observations with unusually high leverage?

```r
par(mfrow = c(2, 2))
plot(lm.fit9)
```



As in question 8, the plot of residuals versus fitted values indicates non linearity in the data. The plot of standardized residuals versus leverage shows there are some outliers and one leverage point (14).

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

```
lm.fit9e = lm(mpg~cylinders*displacement+displacement*weight, data= Auto_new)
summary(lm.fit9e)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto_new)
##
## 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
```

From the p-values, we can see that the interaction between displacement and weight is statistically signifcant.

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

```
lm.fit9f = lm(mpg ~ displacement+log(horsepower)+ log(weight)+sqrt(acceleration)+year+origin, data = Au
summary(lm.fit9f)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + log(horsepower) + log(weight) +
##     sqrt(acceleration) + year + origin, data = Auto_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0441 -1.9150 -0.0836  1.6173 12.5383
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        132.571959  10.833652  12.237  < 2e-16 ***
## displacement         0.013652   0.004553   2.999  0.00289 **
## log(horsepower)     -6.652262   1.549812  -4.292 2.24e-05 ***
## log(weight)        -16.941709   1.964743  -8.623  < 2e-16 ***
## sqrt(acceleration)  -1.306802   0.817670  -1.598  0.11082
## year                 0.749949   0.046770  16.035  < 2e-16 ***
## origin               1.099033   0.250436   4.388 1.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.031 on 385 degrees of freedom
## Multiple R-squared:  0.8515, Adjusted R-squared:  0.8491
## F-statistic: 367.8 on 6 and 385 DF,  p-value: < 2.2e-16
```

I try some different transformations of the variables that *log(horsepower),log(weight), sqrt(acceleration)*.The result shows Adjusted R-squared increases and only *sqrt(acceleration^2)* variable is not statistically significant.

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.

```
data(Carseats)
lm.fit10 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(lm.fit10)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## 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
```

```
summary(Carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75
##  Median :272.0   Median :117.0   Medium:219   Median :54.50
##  Mean   :264.8   Mean   :115.8                Mean   :53.32
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00
##  Max.   :509.0   Max.   :191.0                Max.   :80.00
##    Education    Urban       US
##  Min.   :10.0   No :118   No :142
##  1st Qu.:12.0   Yes:282   Yes:258
##  Median :14.0
##  Mean   :13.9
##  3rd Qu.:16.0
##  Max.   :18.0
```

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!
Based on the coefficients in the model:
Price: Sales decreases/increases 0.054 when Price increases/decreases 1, given other variables constant.
Urban: If the store is in urban area versus not urban area, it would decrease Sales by 0.022, given other variables constant.
US: If the store is in US, it would increase Sales by 1.200, given other variables constant.

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

$$Sales = 13.043 - 0.054Price - 0.022Urban + 1.200US$$

with $Urban = 1$ if the store is in an urban area and 0 if not, and $US = 1$ if the store is in the US and 0 if not.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?
Based on the p-value for each predictors, we can reject the null hypothesis of Price and US.

(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.

```r
lm.fit10e = lm(Sales ~ Price + US, data = Carseats)
summary(lm.fit10e)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## 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?
Based on the RSE and R^2 of the two linear regressions, they both fit the data similarly, while R^2 for the linear regression from (e) (23.54%) is marginally better than for the linear regression from (d).

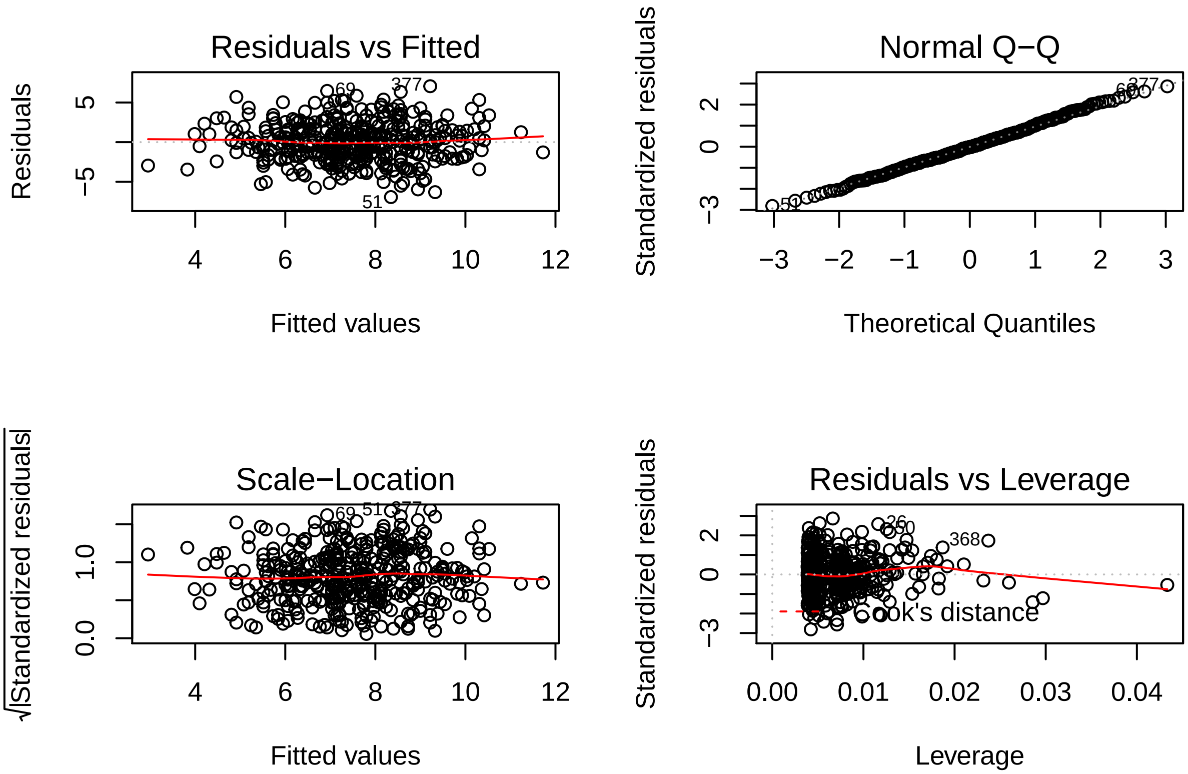
(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```r
confint(lm.fit10e)
```

```
##                   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)?

```r
par(mfrow=c(2,2))
plot(lm.fit10e)
```



There are some outliers (standardized residuals > 2 or < -2)) and some leverage points as some points exceed (p+1)/n(0.0076).