Solutions of the exercises from Chapter 3

Conceptual

Q1. 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.

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

Q2. Carefully explain the differences between the KNN classifier and KNN regression methods.

The KNN classifier is typically used to solve classification problems (those with a qualitative response) by identifying the neighborhood of x_0 *and then estimating the conditional probability* $P(Y = j|X = x_0)$ *for class* j *as the fraction of points in the neighborhood whose response values equal* j. *The KNN regression method is used to solve regression problems (those with a quantitative response) by again identifying the neighborhood of* x_0 *and then estimating* $f(x_0)$ *as the average of all the training responses in the neighborhood.*

Q3. Suppose we have a data set with five predictors, $X1 =$ GPA, $X2 =$ IQ, $X3 =$ Gender (1 for Female and 0 for Male), $X4 =$ Interaction between GPA and IQ, and $X5 =$ 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.

The least square line is given by

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

which becomes for the males

$$\hat{y} = 50 + 20GPA + 0.07IQ + 0.01GPA \times IQ,$$

and for the females

$$\hat{y} = 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ.$$

So the starting salary for males is higher than for females on average iff $50 + 20GPA \geq 85 + 10GPA$ *which is equivalent to* $GPA \geq 3.5$. *Therefore iii. is the right answer.*

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

It suffices to plug in the given values in the least square line for females given above and we obtain

$$\hat{y} = 85 + 40 + 7.7 + 4.4 = 137.1,$$

which gives us a starting salary of 137100$.

(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.*

Q4. 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. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$. 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 knowing more details about the training data, we are not able to know which training RSS is lower between linear or cubic. Essentialy we are using a wrong model, so our results may show some inconsistence.

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

In this case the test RSS depends upon the test data, so we have not enough information to conclude.

(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.

Here also, there is not enough information to tell. If the actual data is pretty linear, then a linear regression may yield better results than a cubic regression. However, if the actual data is more cubic, then cubic regression may be the better option. Whichever one is the better option will yield lower RSS than the other.

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

See answer for (c). above.

Q5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the i-th fitted value takes the form $\hat{y}_i = x_i \hat{\beta}$, where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{k=1}^{n} x_k^2}.$$

Show that we can write

$$\hat{y}_i = \sum_{j=1}^{n} a_j y_j.$$

What is a_j ?

We have immediately that

$$\hat{y}_i = x_i \frac{\sum_{j=1}^n x_j y_j}{\sum_{k=1}^n x_k^2} = \sum_{j=1}^n \frac{x_i x_j}{\sum_{k=1}^n x_k^2} y_j = \sum_{j=1}^n a_j y_j.$$

Q6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\overline{x}, \overline{y})$.

The least square line equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$, so if we substitute $\overline{x}$ for x we obtain

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

We may conclude that the least square line passes through the point $(\overline{x}, \overline{y})$.

Q7. 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 $\overline{x} = \overline{y} = 0$.

We have the following equalities

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_j y_j^2};$$

with $\hat{y}_i = \hat{\beta}_1 x_i$ we may write

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

and finally

$$R^2 = \frac{2(\sum_i x_i y_i)^2 / \sum_j x_j^2 - (\sum_i x_i y_i)^2 / \sum_j x_j^2}{\sum_j y_j^2} = \frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2 \sum_j y_j^2} = Cor(X, Y)^2.$$

Applied

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. For example :

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

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

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

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.571  -3.259  -0.344   2.763  16.924
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.93586    0.71750    55.7   <2e-16 ***
## horsepower  -0.15784    0.00645   -24.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.91 on 390 degrees of freedom
## Multiple R-squared:  0.606,  Adjusted R-squared:  0.605
## F-statistic:  600 on 1 and 390 DF,  p-value: <2e-16
```

We can answer this question by testing the hypothesis $H_0 : \beta_i = 0 \forall i$. *The p-value corresponding to the F-statistic is 7.032 × 10-81, this indicates a clear evidence of a relationship between "mpg" and "horsepower".*

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

We may note that as the R^2 *is equal to 0.6059, almost 60.5948% of the variability in "mpg" can be explained using "horsepower".*

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

As the coeficient of "horsepower" is negative, the relationship is also negative.

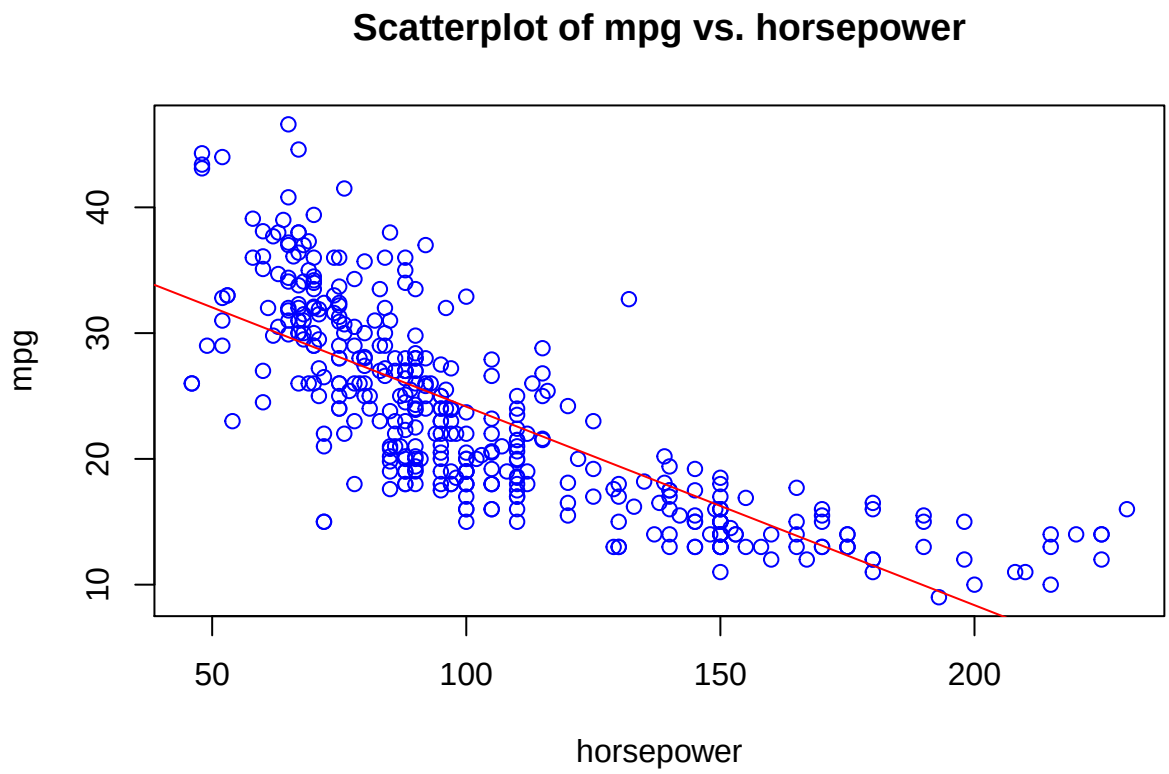
iv. What is the predicted *mpg* associated with a "horsepower" of 98 ? What are the associated 95% confidence and prediction intervals ?

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

```
##     fit   lwr   upr
## 1 24.47 23.97 24.96
```

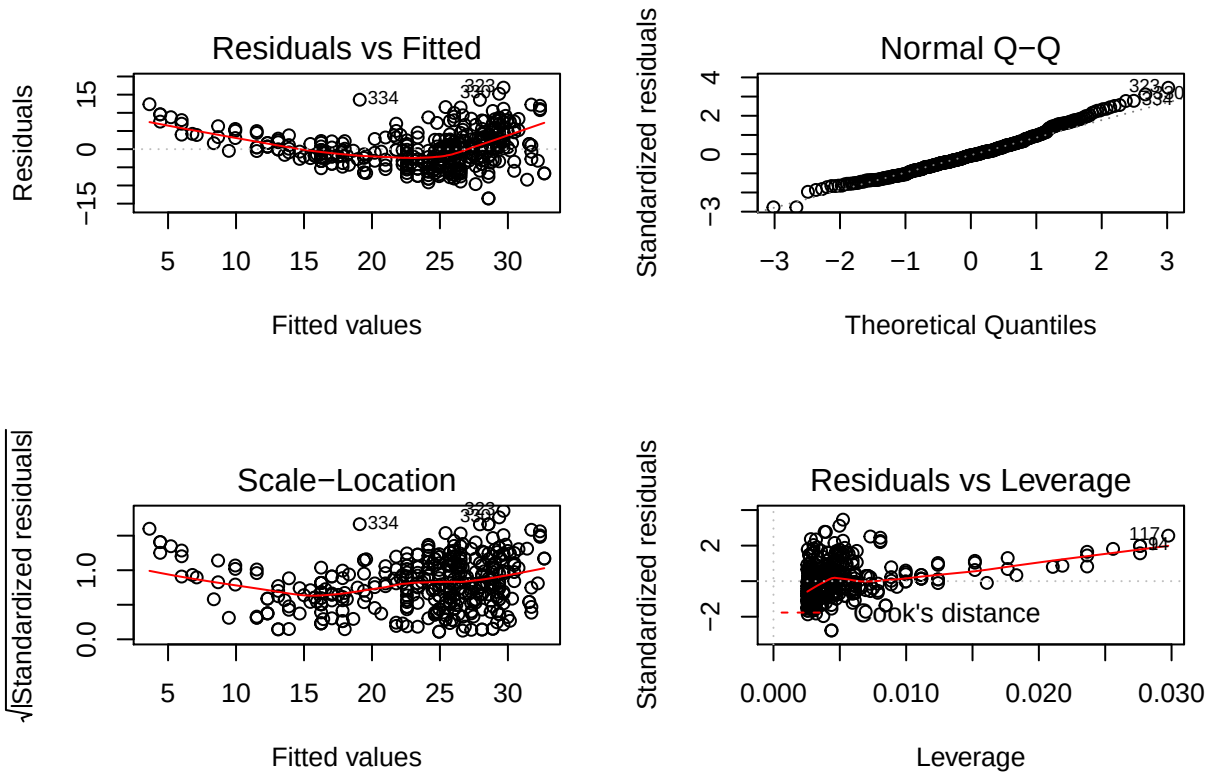
(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

```
plot(Auto$horsepower, Auto$mpg, main = "Scatterplot of mpg vs. horsepower", xlab = "horsepower", ylab =
abline(fit, 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.

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

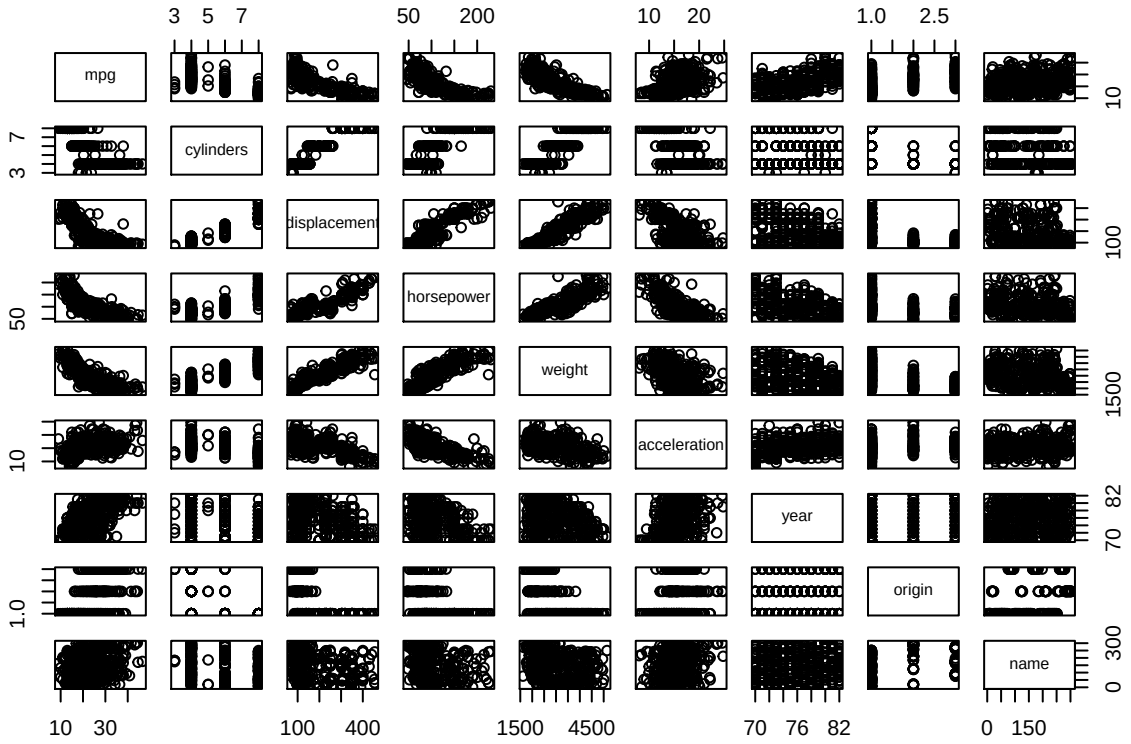


The plot of residuals vs fitted values indicates the presence of non linearity in the data. The plot of standardized residuals vs leverage indicates the presence of a few outliers (higher than 2) and a few high leverage points.

Q9. 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.

```
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.

```
names(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

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

```
##                 mpg cylinders displacement horsepower  weight
## mpg          1.0000   -0.7776      -0.8051    -0.7784 -0.8322
## cylinders   -0.7776    1.0000       0.9508     0.8430  0.8975
## displacement -0.8051    0.9508       1.0000     0.8973  0.9330
## horsepower  -0.7784    0.8430       0.8973     1.0000  0.8645
## weight      -0.8322    0.8975       0.9330     0.8645  1.0000
## acceleration  0.4233   -0.5047      -0.5438    -0.6892 -0.4168
## year          0.5805   -0.3456      -0.3699    -0.4164 -0.3091
## origin        0.5652   -0.5689      -0.6145    -0.4552 -0.5850
##              acceleration    year  origin
## mpg                0.4233  0.5805  0.5652
## cylinders         -0.5047 -0.3456 -0.5689
## displacement      -0.5438 -0.3699 -0.6145
```

```
## horsepower        -0.6892 -0.4164 -0.4552
## weight            -0.4168 -0.3091 -0.5850
## acceleration       1.0000  0.2903  0.2127
## year               0.2903  1.0000  0.1815
## origin             0.2127  0.1815  1.0000
```

(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 ?

```
fit2 <- lm(mpg ~ . - name, data = Auto)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.590 -2.157 -0.117  1.869 13.060
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.72e+01   4.64e+00   -3.71  0.00024 ***
## cylinders    -4.93e-01   3.23e-01   -1.53  0.12780
## displacement  1.99e-02   7.51e-03    2.65  0.00844 **
## horsepower   -1.70e-02   1.38e-02   -1.23  0.21963
## weight       -6.47e-03   6.52e-04   -9.93  < 2e-16 ***
## acceleration  8.06e-02   9.88e-02    0.82  0.41548
## year          7.51e-01   5.10e-02   14.73  < 2e-16 ***
## origin        1.43e+00   2.78e-01    5.13  4.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.33 on 384 degrees of freedom
## Multiple R-squared:  0.821,  Adjusted R-squared:  0.818
## F-statistic:  252 on 7 and 384 DF,  p-value: <2e-16
```

We can answer this question by again testing the hypothesis $H_0 : \beta_i = 0 \forall i$. The p-value corresponding to the F-statistic is 2.0371 × 10-139, 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 ?

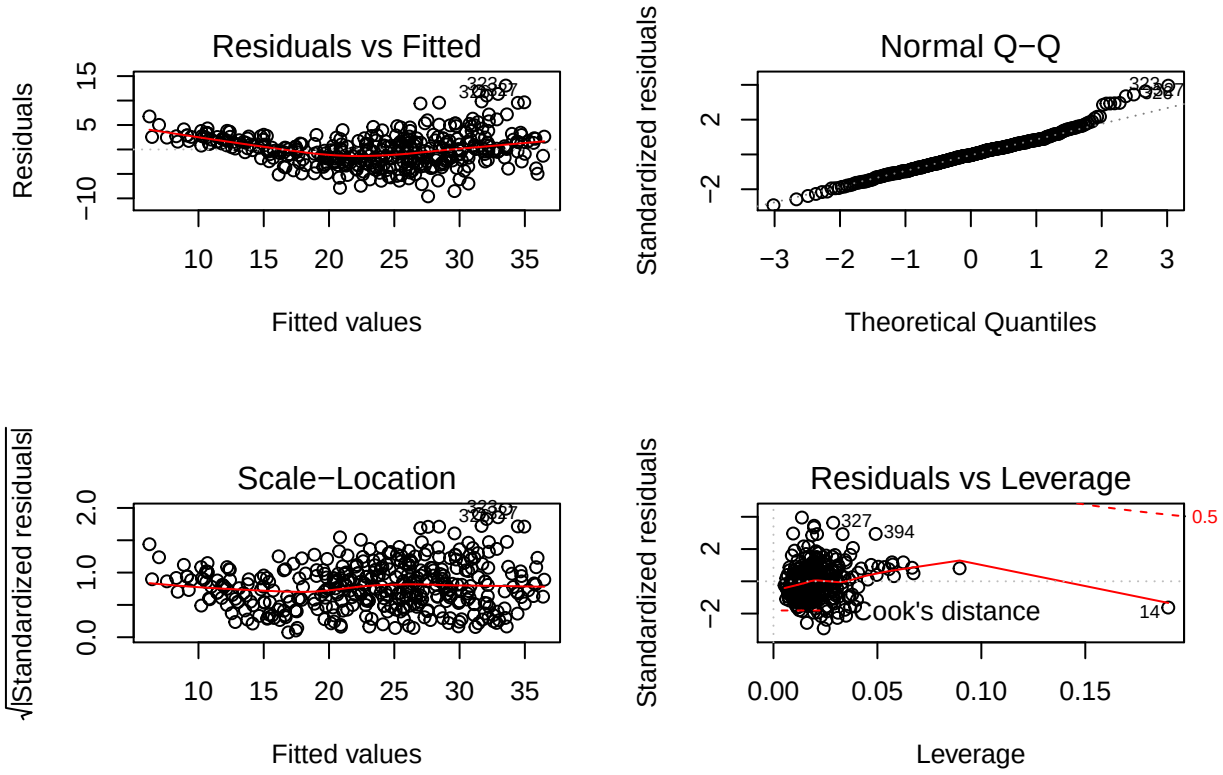
We can answer this question by checking the p-values associated with each predictor's t-statistic. We may conclude that all predictors are statistically significant except "cylinders", "horsepower" and "acceleration".

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 0.7508 in "mpg" (all other predictors remaining constant).

(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 ?

```
par(mfrow = c(2, 2))
plot(fit2)
```



As before, the plot of residuals vs fitted values indicates the presence of mild non linearity in the data. The plot of standardized residuals vs leverage indicates the presence of a few outliers (higher than 2) and a few high leverage points.

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

```
fit3 <- lm(mpg ~ .*., data = Auto[, 1:8])
summary(fit3)
```

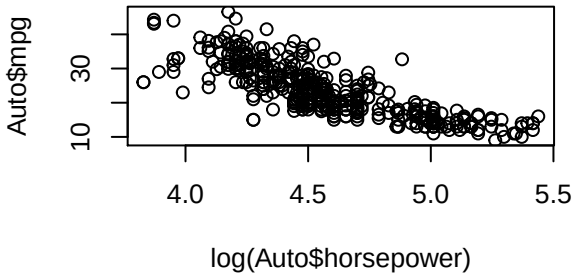
```
##
## Call:
## lm(formula = mpg ~ . * ., data = Auto[, 1:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -7.63  -1.45    0.06    1.27   11.14
##
```

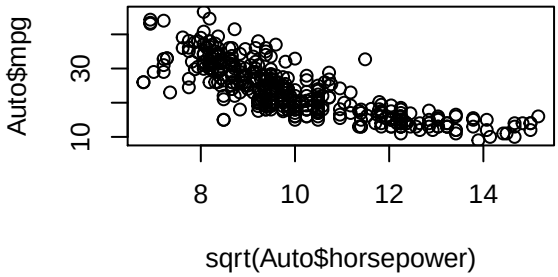
```
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.55e+01   5.31e+01    0.67   0.5047
## cylinders                    6.99e+00   8.25e+00    0.85   0.3974
## displacement                -4.79e-01   1.89e-01   -2.53   0.0119 *
## horsepower                   5.03e-01   3.47e-01    1.45   0.1477
## weight                       4.13e-03   1.76e-02    0.23   0.8144
## acceleration                -5.86e+00   2.17e+00   -2.70   0.0074 **
## year                         6.97e-01   6.10e-01    1.14   0.2534
## origin                      -2.09e+01   7.10e+00   -2.94   0.0034 **
## cylinders:displacement      -3.38e-03   6.46e-03   -0.52   0.6005
## cylinders:horsepower         1.16e-02   2.42e-02    0.48   0.6316
## cylinders:weight             3.58e-04   8.96e-04    0.40   0.6900
## cylinders:acceleration       2.78e-01   1.66e-01    1.67   0.0958 .
## cylinders:year              -1.74e-01   9.71e-02   -1.79   0.0739 .
## cylinders:origin             4.02e-01   4.93e-01    0.82   0.4148
## displacement:horsepower     -8.49e-05   2.89e-04   -0.29   0.7687
## displacement:weight          2.47e-05   1.47e-05    1.68   0.0934 .
## displacement:acceleration   -3.48e-03   3.34e-03   -1.04   0.2985
## displacement:year            5.93e-03   2.39e-03    2.48   0.0135 *
## displacement:origin          2.40e-02   1.95e-02    1.23   0.2187
## horsepower:weight           -1.97e-05   2.92e-05   -0.67   0.5012
## horsepower:acceleration     -7.21e-03   3.72e-03   -1.94   0.0533 .
## horsepower:year             -5.84e-03   3.94e-03   -1.48   0.1392
## horsepower:origin            2.23e-03   2.93e-02    0.08   0.9393
## weight:acceleration          2.35e-04   2.29e-04    1.03   0.3060
## weight:year                 -2.25e-04   2.13e-04   -1.06   0.2918
## weight:origin               -5.79e-04   1.59e-03   -0.36   0.7162
## acceleration:year            5.56e-02   2.56e-02    2.17   0.0303 *
## acceleration:origin          4.58e-01   1.57e-01    2.93   0.0037 **
## year:origin                  1.39e-01   7.40e-02    1.88   0.0606 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 363 degrees of freedom
## Multiple R-squared:  0.889,  Adjusted R-squared:  0.881
## F-statistic:  104 on 28 and 363 DF,  p-value: <2e-16
```

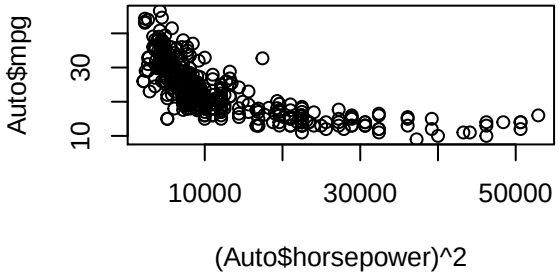
The interactions between "acceleration:origin", "acceleration:year" and "displacement:year" appear to be statistically significant.

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

```
par(mfrow = c(2, 2))
plot(log(Auto$horsepower), Auto$mpg)
plot(sqrt(Auto$horsepower), Auto$mpg)
plot((Auto$horsepower)^2, Auto$mpg)
```







We limit ourselves to examining "horsepower" as sole predictor. It seems that the log transformation gives the most linear looking plot.

Q10. 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)
fit3 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit3)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.921 -1.622 -0.056  1.579  7.058
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.04347    0.65101   20.04  < 2e-16 ***
## Price       -0.05446    0.00524  -10.39  < 2e-16 ***
## UrbanYes    -0.02192    0.27165   -0.08     0.94
## USYes        1.20057    0.25904    4.63  4.9e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.47 on 396 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.234
## F-statistic: 41.5 on 3 and 396 DF,  p-value: <2e-16
```

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

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 54.4588 units in sales. The coefficient of the "Urban" variable may be interpreted by saying that on average the unit sales in urban location are 21.9162 units less than in rural location. The coefficient of the "US" variable may be interpreted by saying that on average the unit sales in a US store are 1200.5727 units more than in a non US store.

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

The model may be written as

$$Sales = 13.0435 + (-0.0545) \times Price + (-0.0219) \times Urban + (1.2006) \times US + \varepsilon$$

with $Urban = 1$ *if the store is in an urban location 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 $H_0 : \beta_j = 0$?

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.

```
fit4 <- lm(Sales ~ Price + US, data = Carseats)
summary(fit4)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.927 -1.629 -0.057  1.577  7.052
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098   20.65  < 2e-16 ***
## Price       -0.05448    0.00523  -10.42  < 2e-16 ***
## USYes        1.19964    0.25846    4.64  4.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.47 on 397 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.235
## F-statistic: 62.4 on 2 and 397 DF,  p-value: <2e-16
```

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

The R^2 for the smaller model is marginally better than for the bigger model. Essentially about 23.9263% of the variability is explained by the model.

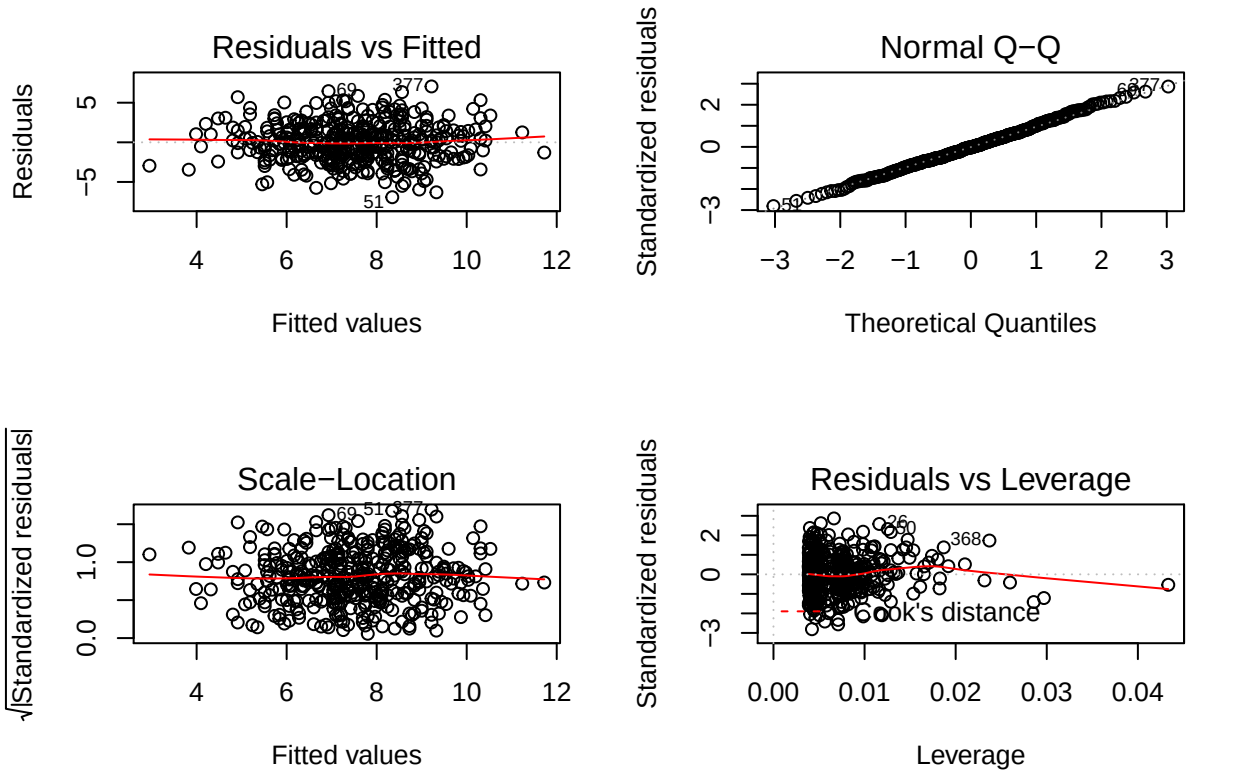
(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
confint(fit4)
```

```
##                2.5 %  97.5 %
## (Intercept) 11.79032 14.2713
## Price       -0.06476 -0.0442
## USYes        0.69152  1.7078
```

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

```
par(mfrow = c(2, 2))
plot(fit4)
```



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

Q11. In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
set.seed(1)
x <- rnorm(100)
y <- 2 * x + rnorm(100)
```

(a) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis H_0. Comment on these results.

```
fit5 <- lm(y ~ x + 0)
summary(fit5)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.915 -0.647 -0.177  0.506  2.311
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x    1.994      0.106    18.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.959 on 99 degrees of freedom
## Multiple R-squared:  0.78,   Adjusted R-squared:  0.778
## F-statistic:  351 on 1 and 99 DF,  p-value: <2e-16
```

According to the summary above, we have a value of 1.9939 for $\hat{\beta}$, a value of 0.1065 for the standard error, a value of 18.7259 for the t-statistic and a value of 2.6422 × 10-34 for the p-value. The small p-value allows us to reject H_0.

(b) Now perform a simple linear regression of x onto y, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis H_0. Comment on these results.

```
fit6 <- lm(x ~ y + 0)
summary(fit6)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.870 -0.237  0.103  0.286  0.894
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y   0.3911     0.0209    18.7   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.425 on 99 degrees of freedom
## Multiple R-squared:  0.78,   Adjusted R-squared:  0.778
## F-statistic:  351 on 1 and 99 DF,  p-value: <2e-16
```

According to the summary above, we have a value of 0.3911 for $\hat{\beta}$, a value of 0.0209 for the standard error, a value of 18.7259 for the t-statistic and a value of 2.6422 × 10-34 for the p-value. The small p-value allows us to reject H_0.

(c) What is the relationship between the results obtained in (a) and (b) ?

We obtain the same value for the t-statistic and consequently the same value for the corresponding p-value.

(d) For the regrssion of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $\hat{\beta}/SE(\hat{\beta})$, where $\hat{\beta}$ is given by (3.38), and where

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i=1}^{n} x_i^2}}.$$

Show algebraically, and confirm numerically in R, that the t-statistic can be written as

$$\frac{\sqrt{n-1}\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i y_i)}}.$$

We have

$$t = \frac{\sum_i x_i y_y / \sum_j x_j^2}{\sqrt{\sum_i (y_i - x_i\hat{\beta})^2/(n-1)\sum_j x_j^2}} = \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{\sum_j x_j^2 \sum_i (y_i - x_i \sum_j x_j y_j / \sum_j x_j^2)^2}} = \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{(\sum_j x_j^2)(\sum_j y_j^2) - (\sum_j x_j y_j)}}.$$

Now let's verify this result numerically.

```
n <- length(x)
t <- sqrt(n - 1)*(x %*% y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
as.numeric(t)
```

```
## [1] 18.73
```

We may see that the t above is exactly the t-statistic given in the summary of "fit6".

(e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same t-statistic for the regression of x onto y.

It is easy to see that if we replace x_i by y_i in the formula for the t-statistic, the result would be the same.

(f) In R, show that when regression is performed with an intercept, the t-statistic for $H_0 : \beta_1 = 0$ is the same for the regression of y onto x as it is the regression of x onto y.

```r
fit7 <- lm(y ~ x)
summary(fit7)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.877 -0.614 -0.140  0.539  2.346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0377     0.0970   -0.39      0.7
## x             1.9989     0.1077   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.963 on 98 degrees of freedom
## Multiple R-squared:  0.778,  Adjusted R-squared:  0.776
## F-statistic:  344 on 1 and 98 DF,  p-value: <2e-16
```

```r
fit8 <- lm(x ~ y)
summary(fit8)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9085 -0.2810  0.0627  0.2457  0.8574
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0388     0.0427    0.91     0.37
## y             0.3894     0.0210   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.425 on 98 degrees of freedom
## Multiple R-squared:  0.778,  Adjusted R-squared:  0.776
## F-statistic:  344 on 1 and 98 DF,  p-value: <2e-16
```

It is again easy to see that the t-statistic for "fit7" and "fit8" are both equal to 18.5556.