Chapter - 3

Q8. sThis 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 :

b. 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
## -13.5710  -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
```

We can find an answer to this question by putting the hypothesis H0:i=0 I to the test. The F-p-value statistic's is 7.03198910-81, indicating that there is clear evidence of a relationship between "mpg" and "horsepower."

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

We use the mean of the response and the RSE to calculate the residual error relative to the response. The average mpg is 23.4459184 miles per gallon. The lm. fit RSE was 4.9057569, indicating a percentage error of 20.9237141 percent. It's also worth noting that, with an R2 of 0.6059483, "horsepower" can explain nearly 60.5948258 percent of the variation in "mpg."

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

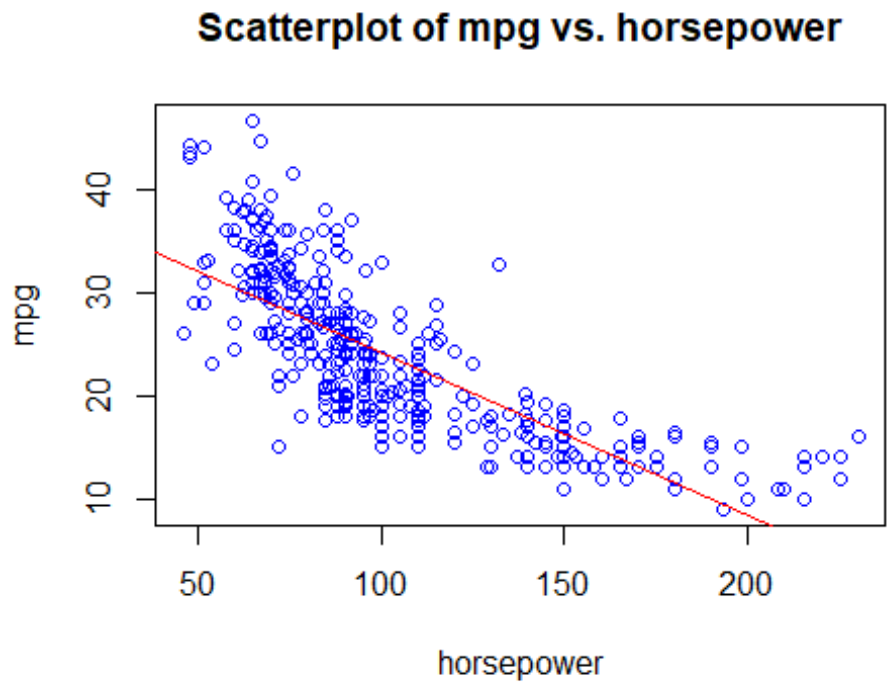
Because the “horsepower” coeficient is negative, the relationship is also negative. The linear regression indicates that the more horsepower an automobile has, the lower the mpg fuel efficiency.

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 24.46708 23.97308 24.96108

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

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

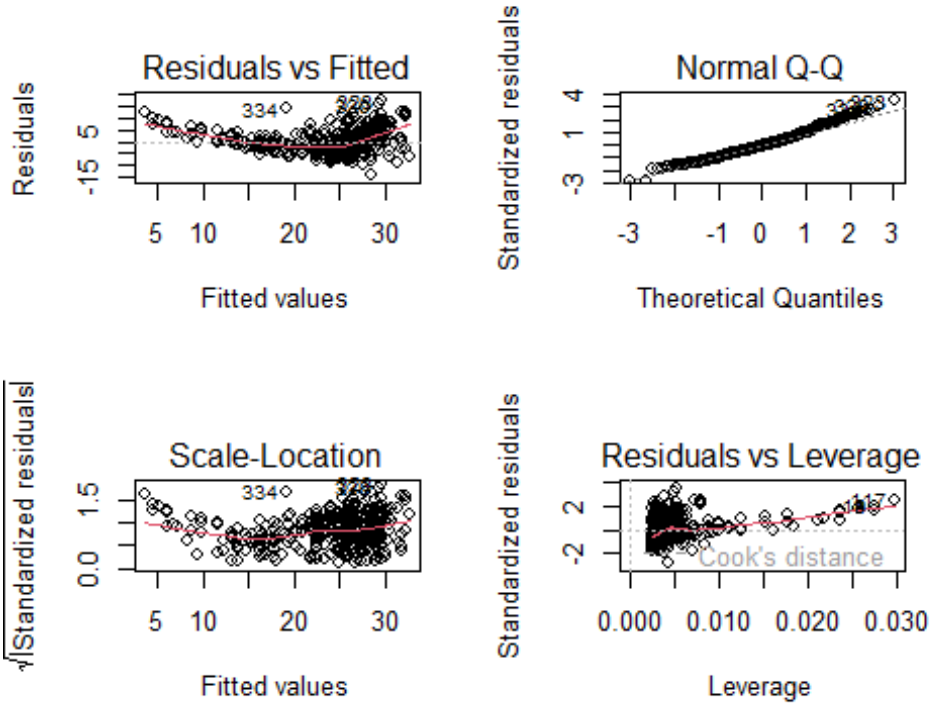
b. Plot the response and the predictor. Use the abline() function to display the least squares regression line.

```
plot(Auto$horsepower, Auto$mpg, main = "Scatterplot of mpg vs. horsepower",
xlab = "horsepower", ylab = "mpg", col = "blue")
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)
```

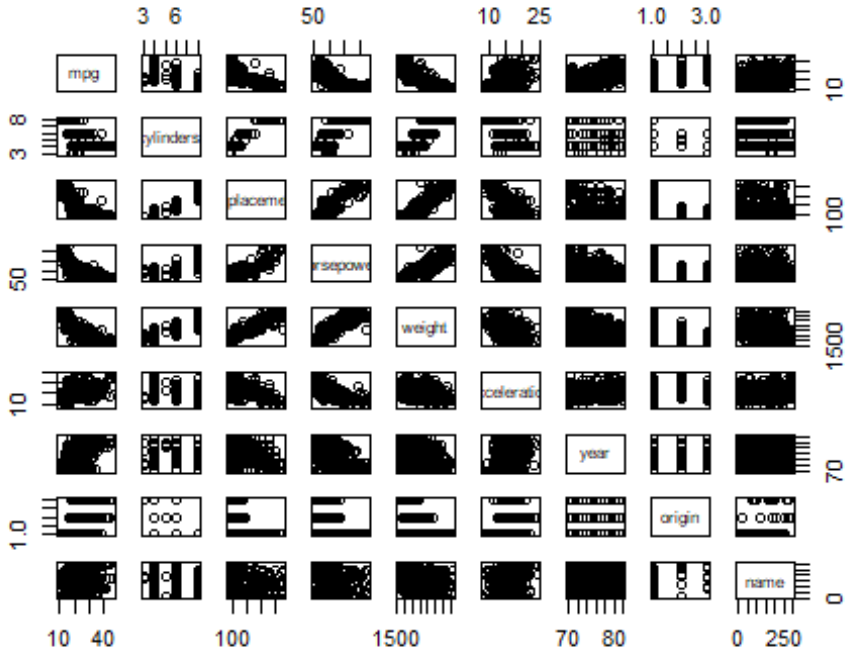


The presence of non linearity in the data is indicated by the plot of residuals versus fitted values. The plot of standardized residuals vs. leverage reveals a few outliers (values greater than 2 or less than -2) as well as a few high leverage points.

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"   "weight"
## [6] "acceleration" "year"         "origin"       "name"

cor(Auto[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 :

d. 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
## -9.5903 -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
```

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

This question can be answered by looking at the p-values associated with each predictor’s t-statistic. Except for “cylinders,” “horsepower,” and “acceleration,” we can conclude that all predictors are statistically significant.

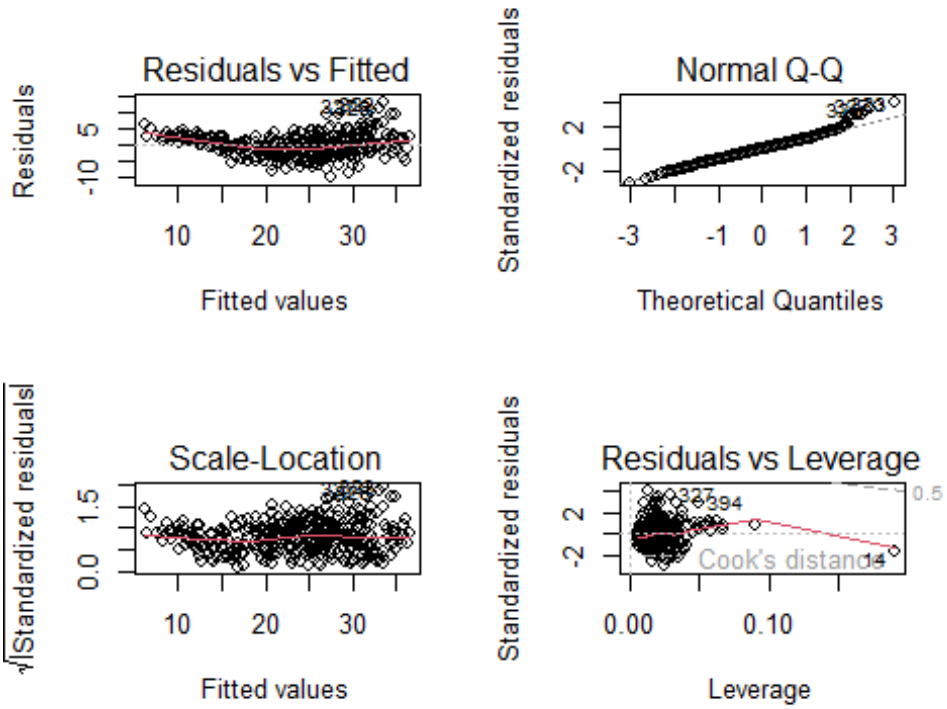
iii. What does the coefficient for the “year” variable suggest ?

The coefficient of the “year” variable indicates that an increase of one year results in an increase of 0.7507727 in “mpg” (all other predictors remaining constant). In other words, cars improve their fuel efficiency by nearly 1 mpg per 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 ?

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



The plot of residuals versus fitted values, as before, indicates that the data has mild non linearity. The plot of standardized residuals versus leverage reveals a few outliers (values greater than 2 or less than -2) as well as 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 ?

We took the two most highly correlated pairs from the correlation matrix and used them to select interaction effects.

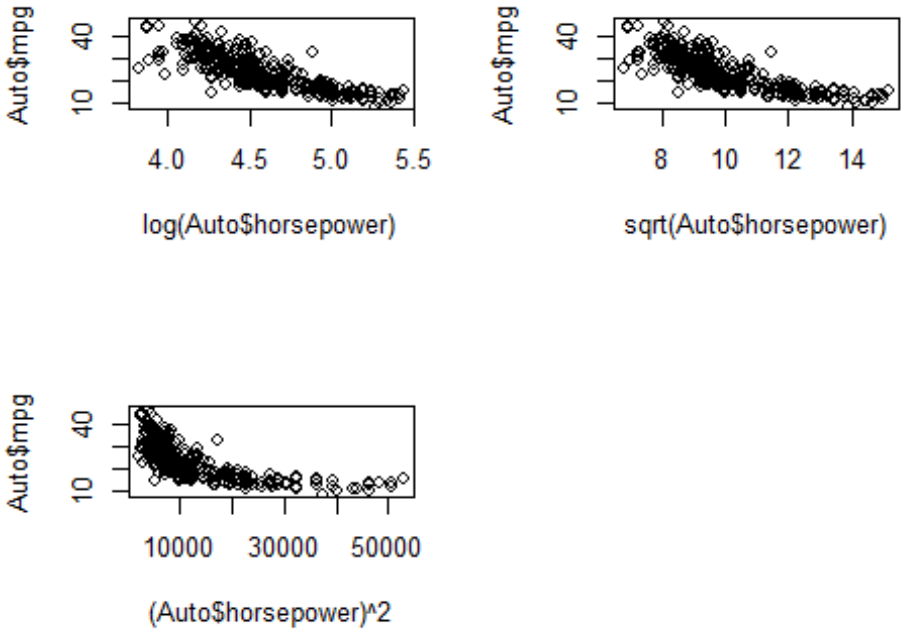
```
fit3 <- lm(mpg ~ cylinders * displacement+displacement * weight, data =
Auto[, 1:8])
summary(fit3)

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

The interaction between displacement and weight is statistically significant, but the interaction between cylinders and displacement is not, as shown by the p-values.

f. Try a few different transformations of the variables, such as logX, X−−√, X2. 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're only looking at "horsepower" as a single predictor. The log transformation appears to produce the most linear plot.