# DSI-06 Homework 2: Chapter 3, pg 123

2023-02-24

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

```
install.packages("ISLR") #install package containing Auto dataset
```

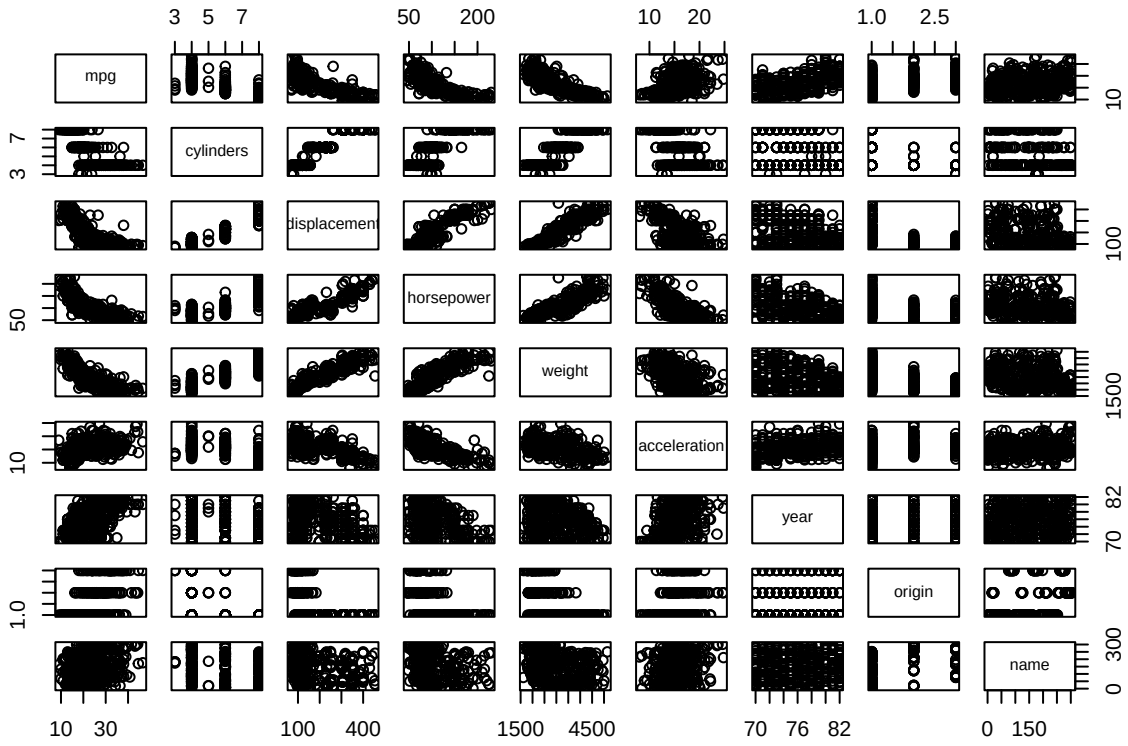
```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(ISLR)
attach(Auto) #attach Auto dataset to make the variables associated with Auto available.
head(Auto) #return the column names and first few rows of the dataset
```

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

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

```
plot(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.**

```
cor(Auto[ ,-9]) #this is saying exclude the 9th column, which is name!
```

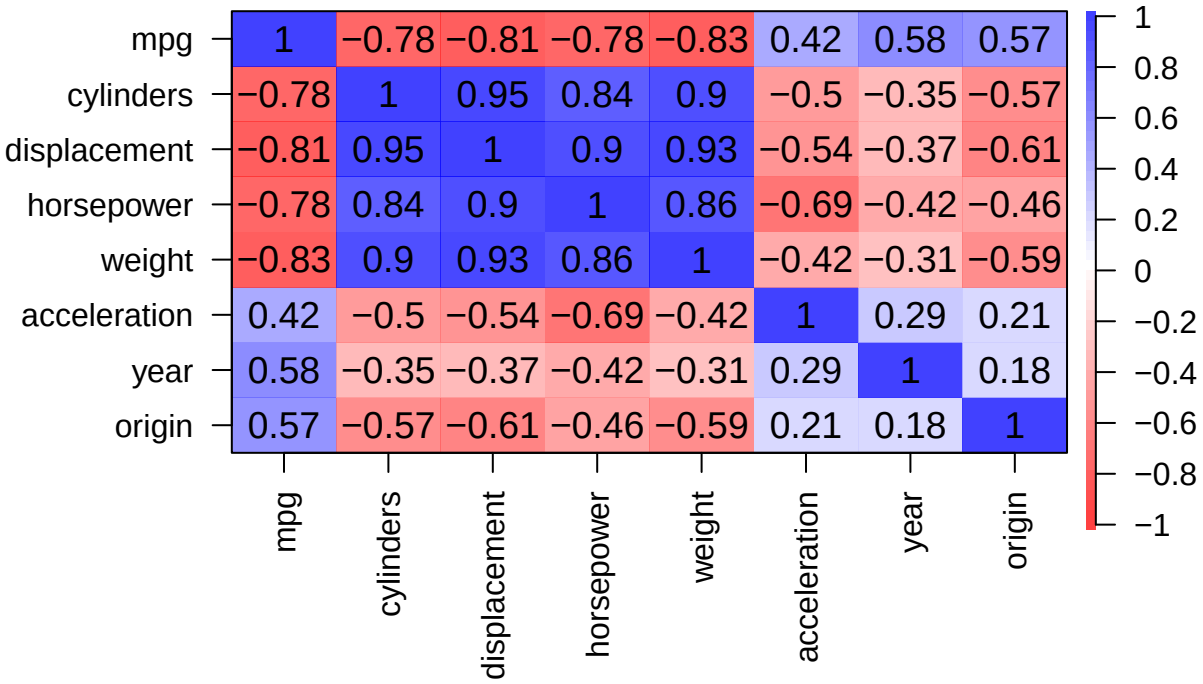
```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -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
```

Alternatively, we can use the psych package to plot the correlations, giving us a nice easy-to-read figure!

```
install.packages("psych")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library("psych")
Auto_cor <- cor(Auto[ ,-9])
cor.plot(Auto_cor, xlas = 2) #xlas =2 rotates the variable labels for better readibility
```




**(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?**

```
Auto_mult_lm <- lm(mpg ~ .-name, data = Auto) #this is a short cut to writing out all the variables min
#alternatively, could write out:
#Auto_mult_lm <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin

summary(Auto_mult_lm)
```

```
##
## 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
```

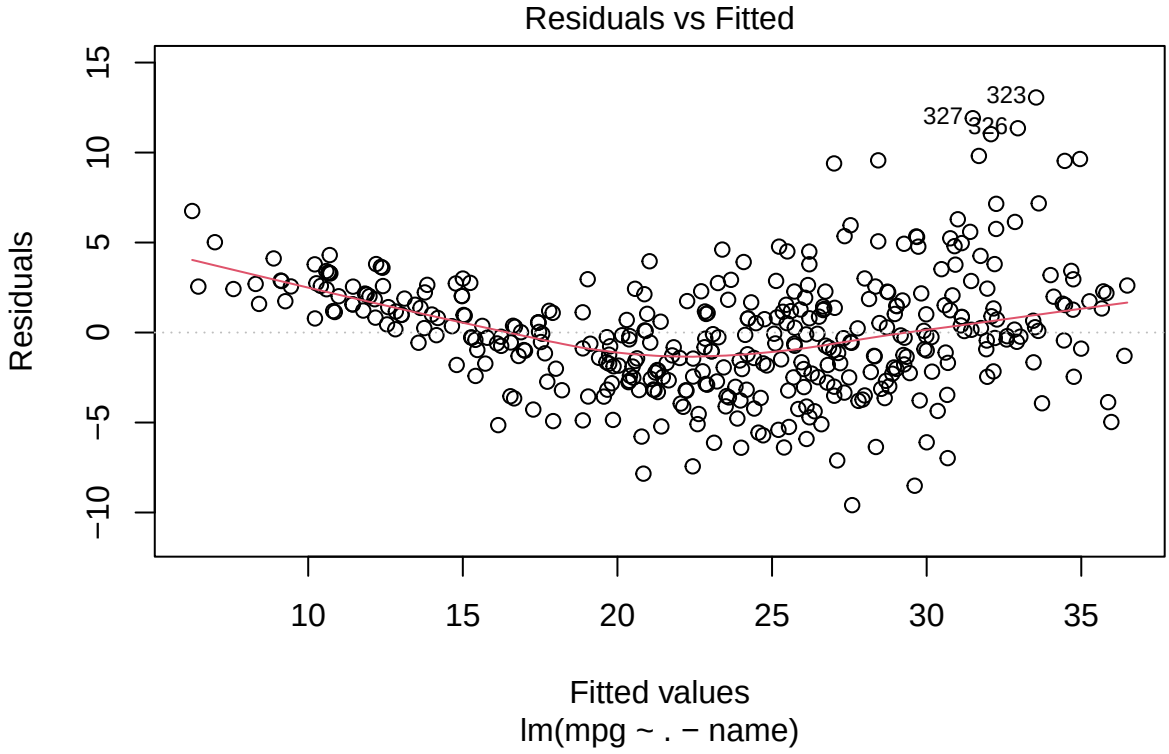
i. We see there is a relationship between the predictors and the response (large F-statistic, p-value for the overall model is significant, as well has large R^2!)

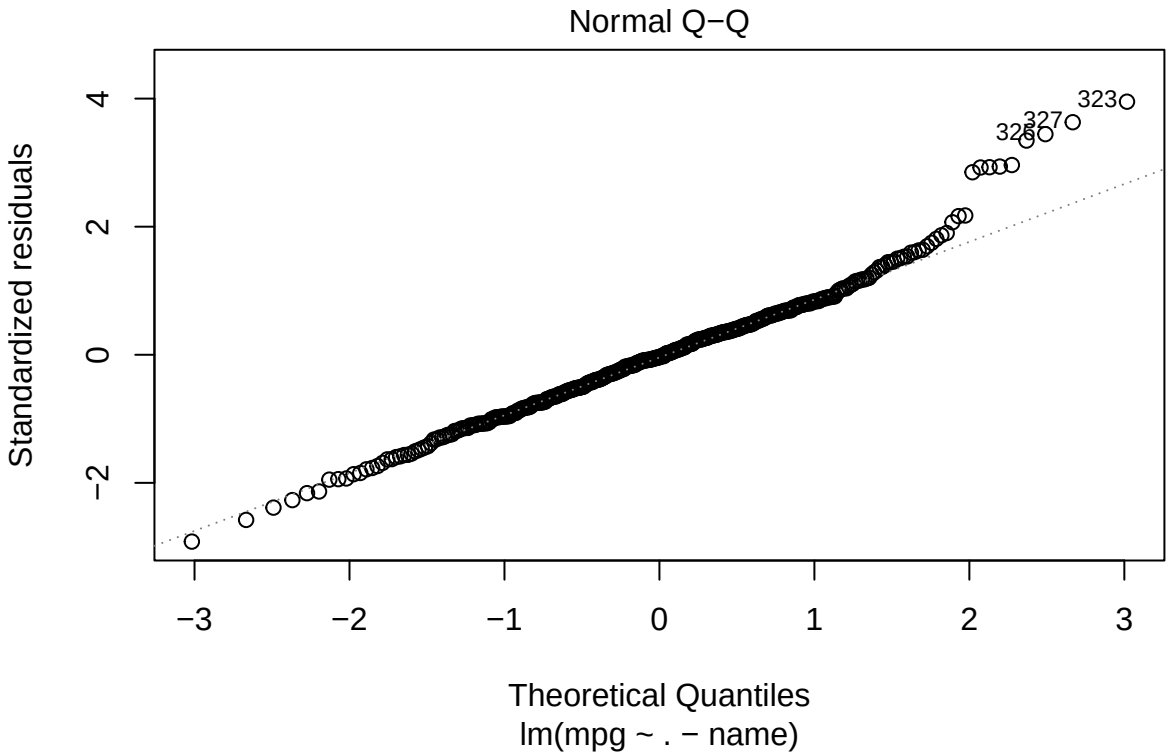
ii. Displacement, weight, year and origin seem to have a statistically significant relationship to mpg (sig p-values)

iii. The coefficient for the year variable suggests, per every unit increase of year there is a corresponding increase in mpg by 0.750773, assuming all other predictors are held 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 plot identify any observations with unusually high leverage?**

```
plot(Auto_mult_lm, which = c(1, 2))
```






There is a slight trend in the residuals illustrated by the red curve which could indicate patterns in the data that are not captured by the linear model!

The Q-Q plot shows some linearity of the data up to the 2nd theoretical quantile, however after that the residuals vs fitted plot seems to deviate from the line!

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

Note, we can look at the correlation plot from part a to find variables that appear to be correlated! i.e,cylinder and displacement have a correlation of 0.95 and weight and displacement have a correlation of 0.93!

```
Auto_mult_lm_int <- lm(mpg ~ .-name + cylinders*displacement + cylinders*weight, data = Auto)
summary(Auto_mult_lm_int)
```

```
##
## Call:
## lm(formula = mpg ~ . - name + cylinders * displacement + cylinders *
##     weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1234  -1.7125  -0.1423   1.4285  12.3588
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.6255027  5.0920727   1.301  0.19399
## cylinders           -4.8224843  0.6438744  -7.490 4.83e-13 ***
## displacement        -0.0008817  0.0228354  -0.039  0.96922
## horsepower          -0.0348918  0.0134317  -2.598  0.00975 **
## weight              -0.0136868  0.0020947  -6.534 2.05e-10 ***
```

```
## acceleration            0.0943521  0.0902262   1.046  0.29635
## year                    0.7788369  0.0465573  16.729  < 2e-16 ***
## origin                  0.7592928  0.2681462   2.832  0.00488 **
## cylinders:displacement  0.0023878  0.0031477   0.759  0.44857
## cylinders:weight        0.0013073  0.0003098   4.219 3.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.024 on 382 degrees of freedom
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.8499
## F-statistic:   247 on 9 and 382 DF,  p-value: < 2.2e-16
```

Here, we see cylinder:weight appears to be a significant interaction!