ii.Exercise3.7.9

Kaarthik Sundaramoorthy, Sahil Shah and Vidhi Shah

6/3/2020

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

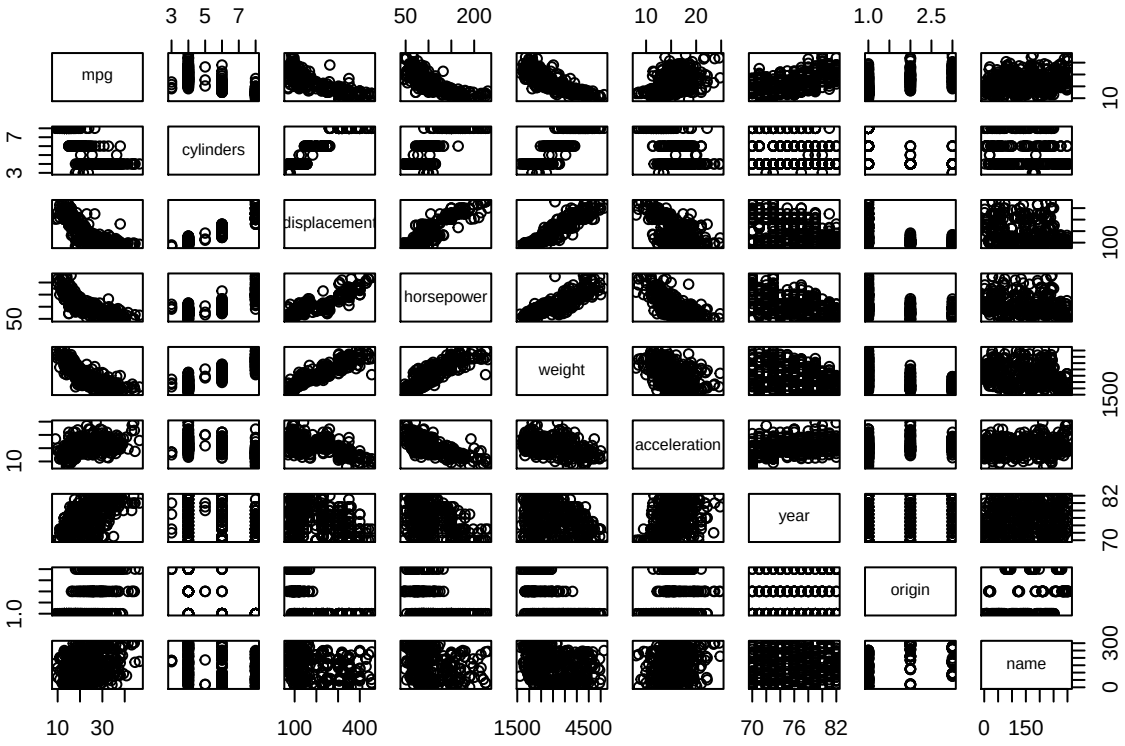
```
require(ISLR)
```

```
## Loading required package: ISLR
```

```
data(Auto)
attach(Auto)
```

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

```
par(mfrow = c(1,1))
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.

```
head(Auto)
```

```
##   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
```

```
##help("cor")
cor(subset(Auto, select = -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
```

(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:

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

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

Ans. Yes, there is a relationship between the predictors and the response variables. The F-Statistics is really far from 1, which results into small p-value. Indicating that there is an evidence against the null hypothesis.

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

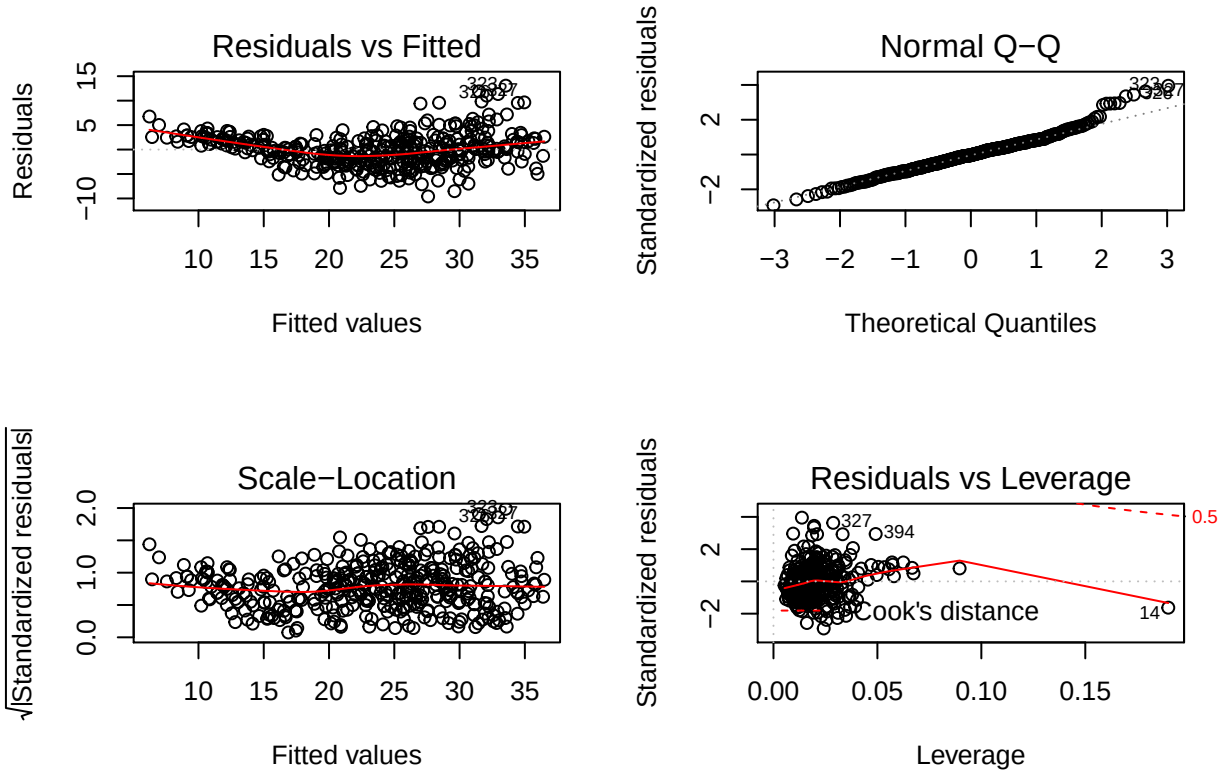
Ans. Predictors like `origin`, `year`, `weight` and `displacement` have the statistically significant relationship with the response variable.

(c) iii. What does the coefficient for the year variable suggest?

Ans. The coefficient for the `year` i.e. 0.75 with the additional year of the car's age, the `mpg` will on average increase by 0.75. In other words it suggests that later model year car will have better `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?

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



Ans. There is an evidence of the non-linearity. The polt of Residuals vs Fitted shows that variance of the error terms increase with the values of the response. The Residuals vs Leverage plot shows no outliers, but the tending of the curve at point 14 shows the sign of high leverage.

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

```
lm.fit1 <- lm(mpg ~ .-name + cylinders:horsepower, data = Auto)
lm.fit2 <- lm(mpg~displacement+origin+year*weight, data = Auto)
lm.fit3 <- lm(mpg~year+origin+displacement*weight, data = Auto)
lm.fit4 <- lm(mpg~cylinders*displacement+displacement*weight, data = Auto)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = mpg ~ . - name + cylinders:horsepower, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2399 -1.6871 -0.0511  1.2858 11.9380
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         11.7025260  4.9115648   2.383 0.017676 *
## cylinders           -4.3060695  0.4580950  -9.400  < 2e-16 ***
```

```
## displacement         -0.0013925  0.0069110  -0.201 0.840426
## horsepower           -0.3156601  0.0306339 -10.304  < 2e-16 ***
## weight               -0.0038948  0.0006231  -6.250 1.09e-09 ***
## acceleration         -0.1703028  0.0901427  -1.889 0.059612 .
## year                  0.7393193  0.0448736  16.476  < 2e-16 ***
## origin                0.9031644  0.2496880   3.617 0.000338 ***
## cylinders:horsepower  0.0402008  0.0037856  10.619  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.929 on 383 degrees of freedom
## Multiple R-squared:  0.8621, Adjusted R-squared:  0.8592
## F-statistic: 299.3 on 8 and 383 DF,  p-value: < 2.2e-16
```

```
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + origin + year * weight, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9402 -1.8736 -0.0966  1.5924 12.2125
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.076e+02  1.290e+01  -8.339 1.34e-15 ***
## displacement -4.020e-04  4.558e-03  -0.088 0.929767
## origin        9.116e-01  2.547e-01   3.579 0.000388 ***
## year          1.962e+00  1.716e-01  11.436  < 2e-16 ***
## weight        2.605e-02  4.552e-03   5.722 2.12e-08 ***
## year:weight  -4.305e-04  5.967e-05  -7.214 2.89e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.145 on 386 degrees of freedom
## Multiple R-squared:  0.8397, Adjusted R-squared:  0.8376
## F-statistic: 404.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = mpg ~ year + origin + displacement * weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6119  -1.7290  -0.0115   1.5609  12.5584
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -8.007e+00  3.798e+00  -2.108   0.0357 *
## year                 8.194e-01  4.518e-02  18.136  < 2e-16 ***
```

```
## origin               3.567e-01  2.574e-01   1.386   0.1666
## displacement        -7.148e-02  9.176e-03  -7.790 6.27e-14 ***
## weight              -1.054e-02  6.530e-04 -16.146  < 2e-16 ***
## displacement:weight  2.104e-05  2.214e-06   9.506  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.016 on 386 degrees of freedom
## Multiple R-squared:  0.8526, Adjusted R-squared:  0.8507
## F-statistic: 446.5 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
summary(lm.fit4)
```

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

Ans. As seen in question (c) part (ii) there are **4** variables that are statistically significant, We are try the significant interactions. There is statistic significance in four of the lm.fit. There is a high correlated pairs in the lm.fit4 looking at the p-value, there is a strong statistical significance between displacement and weight, and there is none between cylinders and displacement.

(f) Try a few different transformations of the variables, such as log(X), $\sqrt{X}$, X2. Comment on your findings.