Linga Sai Yuvesh Venketa, Kotiala

16242113

ISL LAB-2

2. 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?

Ans.

If we don't have the ISLR package installed then install the package using command >install.packages('ISLR')

```
> install.packages('ISLR')
Installing package into 'C:/Users/Yuvesh/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/ISLR_1.2.zip'
Content type 'application/zip' length 2923760 bytes (2.8 MB)
downloaded 2.8 MB

package 'ISLR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Yuvesh\AppData\Local\Temp\RtmpcHPBVR\downloaded_packages
```

After Installing the package, it is to be called by using command

>library(ISLR)

Then the actual relation is between predictor and the response

```
> 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
```

Meaning of lm is linear models, where formula is mpg = horsepower + c

And we see that mpg is depended variable of horsepower

We can answer this question by testing the hypothesis H0:βi=0 ∀i. The p-value corresponding to the F-statistic is 7.03198910^{-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?

The mean can be calculated by the mean function and the relationship can be found out.

```
> mean(Auto$mpg)
[1] 23.44592
```

We get value of fit from the above operation.

To calculate the residual error relative to the response we use the mean of the response and the RSE. The mean of mpg is 23.4459184. The RSE of the lm. fit was 4.9057569 which indicates a percentage error of 20.9237141%. We may also note that as the R2 is equal to 0.6059483, almost 60.5948258% 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 coefficient of "horsepower" is negative, the relationship is also negative. The more horsepower an automobile has the linear regression indicates the less mpg fuel efficiency the automobile will have.

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

Under the interval of confidence:

```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> |
```

Under the interval of prediction:

```
> 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

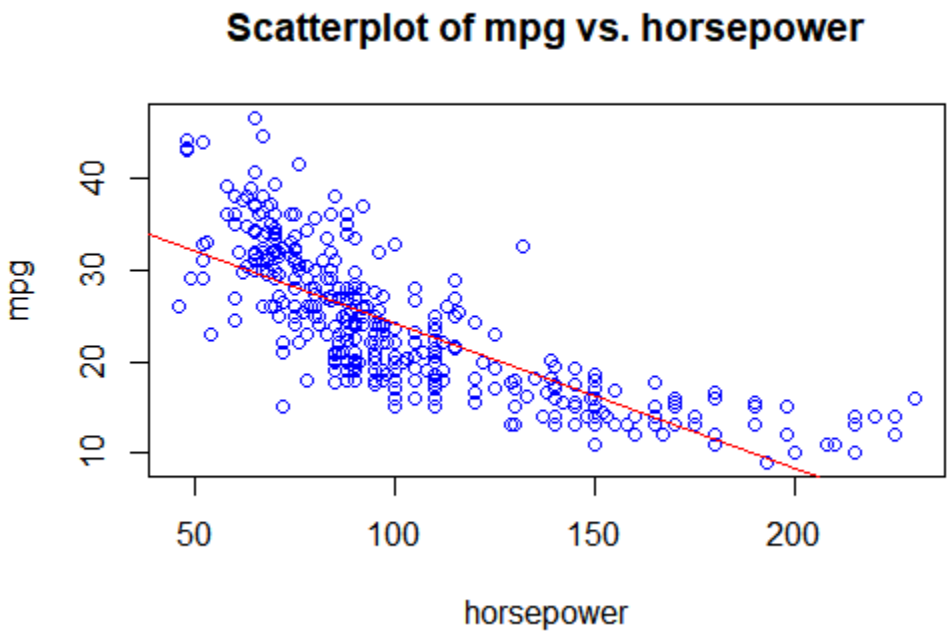
least squares regression line.

Ans:

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

>abline(fit, col = "red")

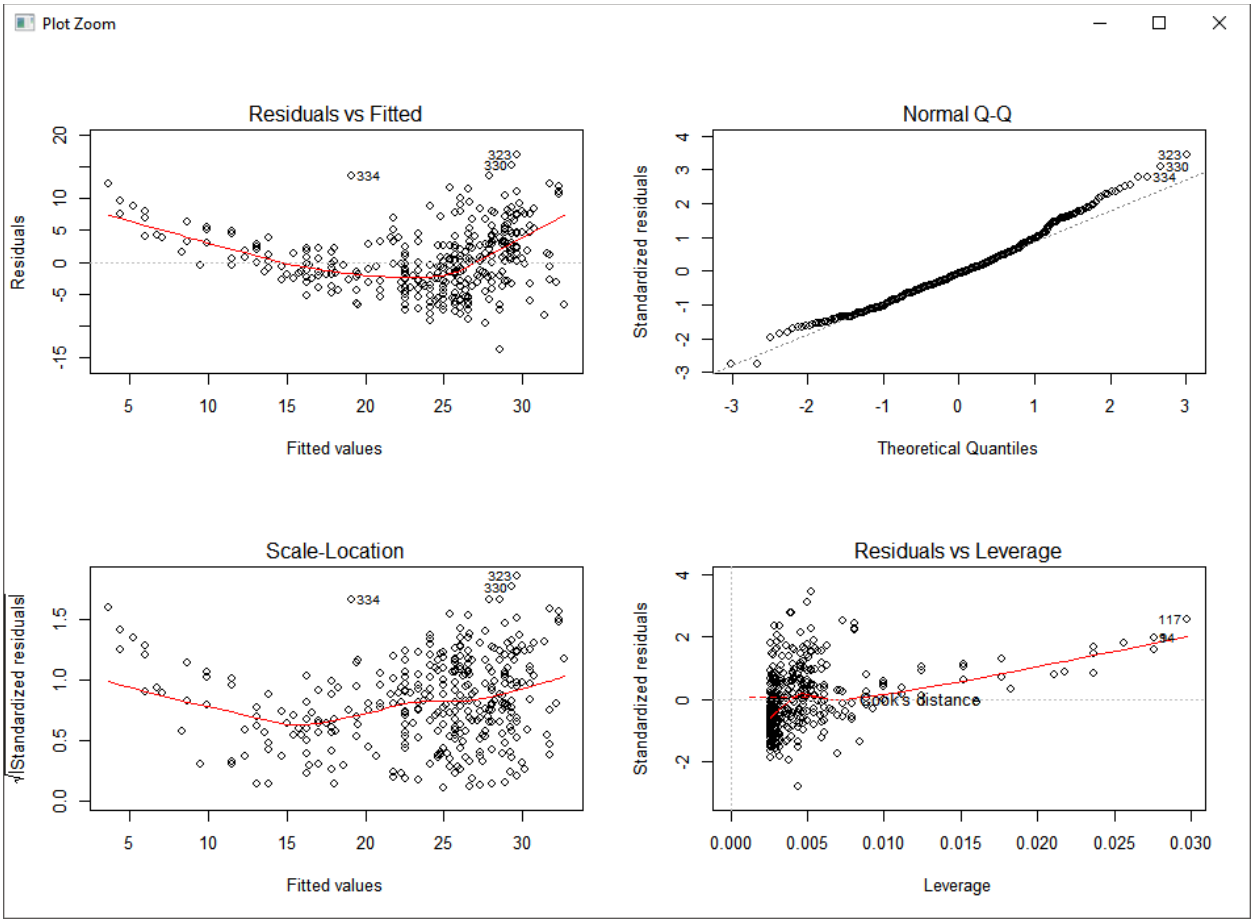
The abline function adds one or more straight lines through the current plot.



(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)



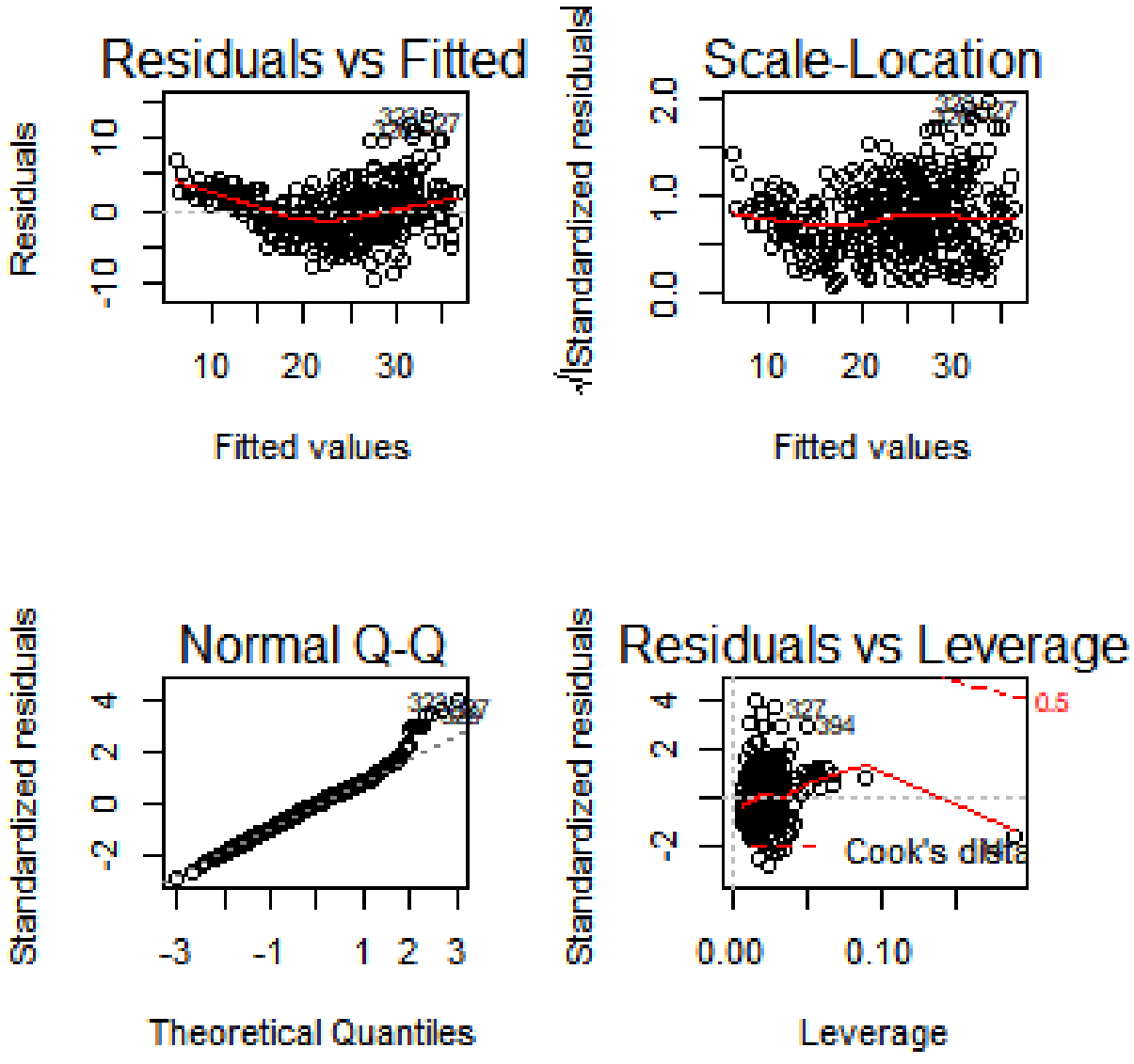
3. This question involves the use of multiple linear regression on the "Auto" data set.

(a)Produce a scatterplot matrix which include all the variables in the data set.

Ans.

Pairs() :A matrix of scatterplots is produced. Which means a of matrix is being scattered on the plot as shown in the figure.

> 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.

Ans.

>cor(Auto[1:8])

```
> cor(Auto[1:8])
                     mpg  cylinders displacement horsepower     weight acceleration       year
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277
                 origin
mpg           0.5652088
cylinders    -0.5689316
displacement -0.6145351
horsepower   -0.4551715
weight       -0.5850054
acceleration  0.2127458
year          0.1815277
origin        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:

```
> 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
```

We can answer this question by again testing the hypothesis H0:βi=0 ∀i. The p-value corresponding to the F-statistic is 2.037105910^{-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?

Ans.

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 ?

Ans.

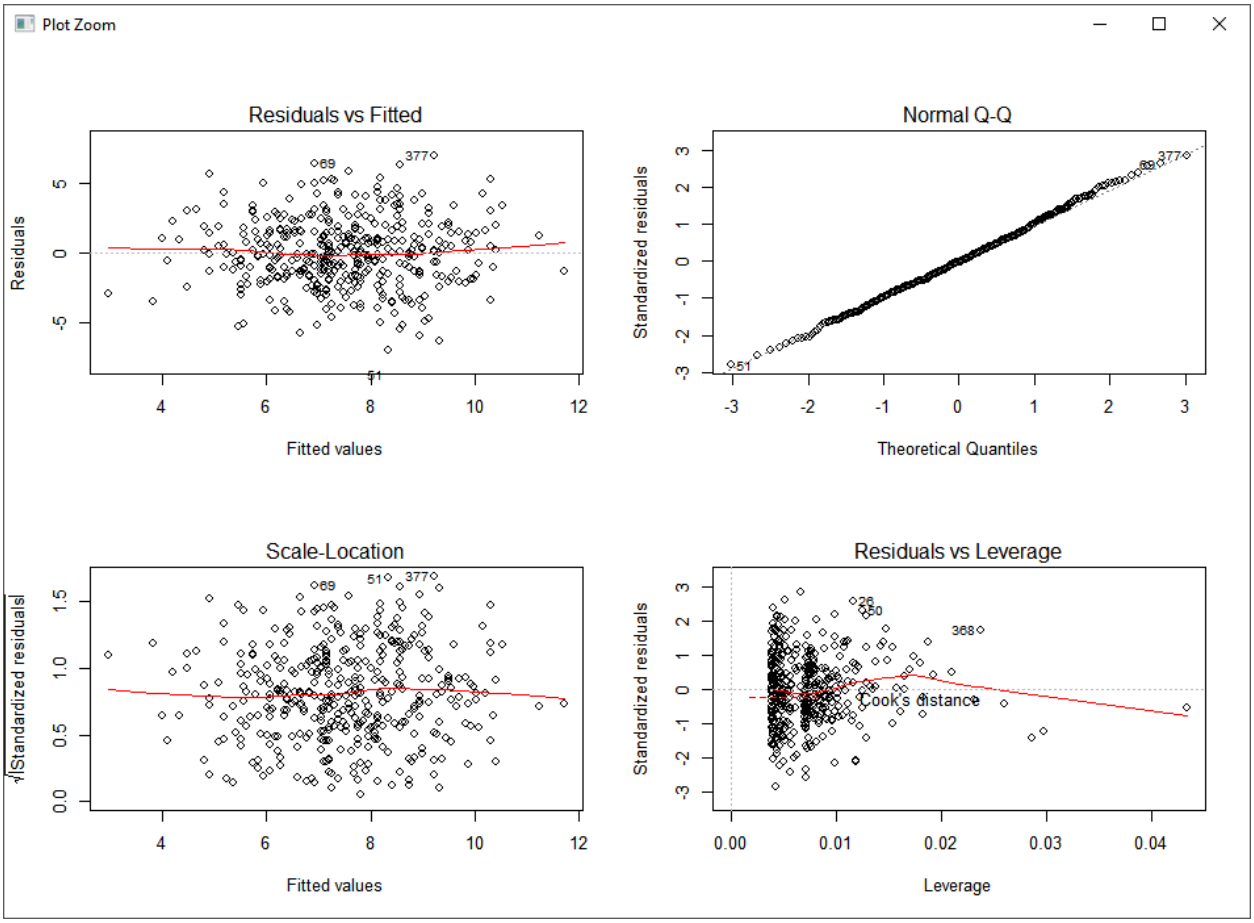
The coefficient ot the "year" variable suggests that the average effect of an increase of 1 year is an increase of 0.7507727 in "mpg" (all other predictors remaining constant). In other words, cars become more fuel efficient every year by almost 1 mpg / year.

(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 ?

Ans.

>par(mfrow = c(2, 2))

>plot(fit2)



As before, the plot of residuals versus fitted values indicates the presence of mild non linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and one high leverage point.

(e) Use the ∗ and : symbols to fit linear regression models with interaction effects.

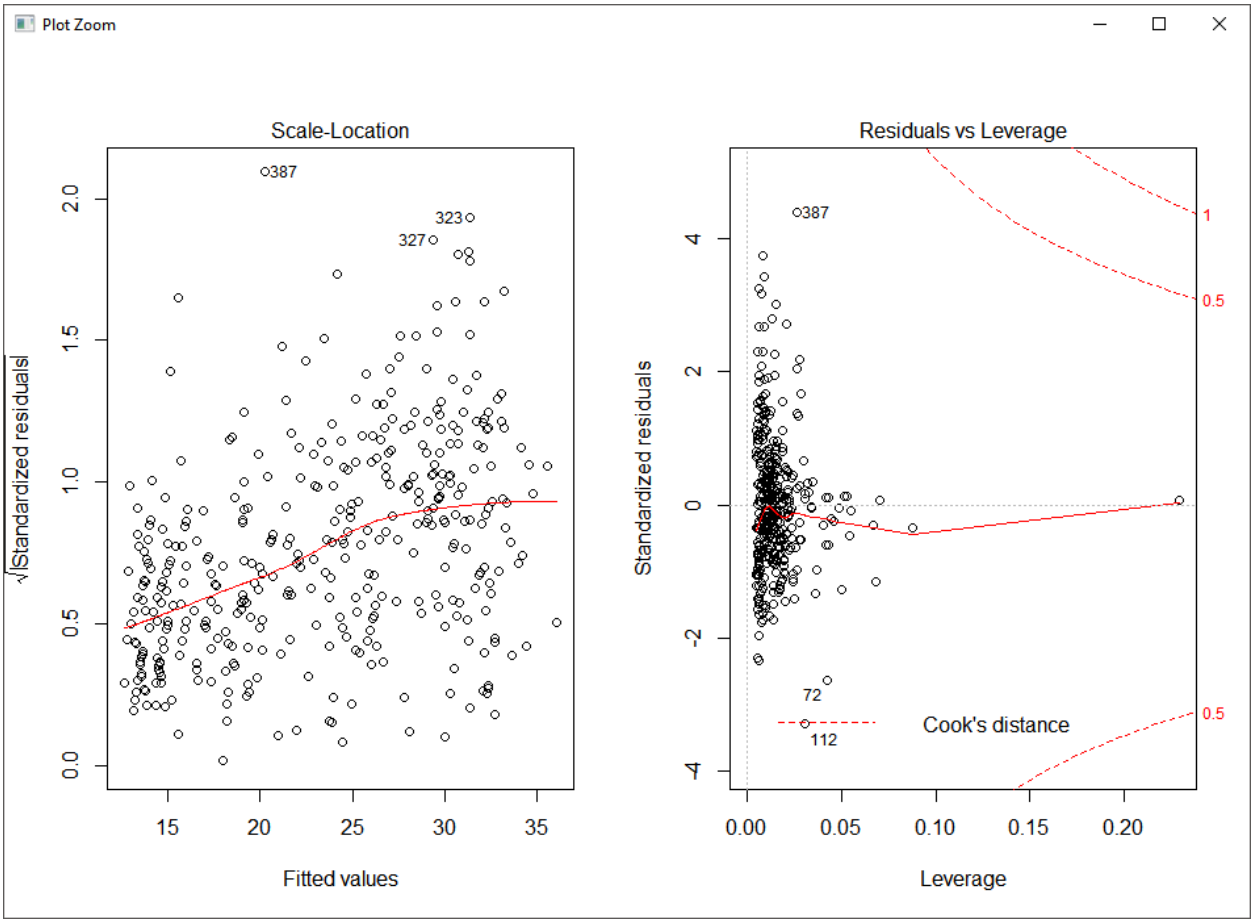
Do any interactions appear to be statistically significant?

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

>summary(fit3)

>par(mfrow = c(1, 2))

>plot(fit3)



(f) Try a few different transformations of the variables, such as log(X), √X, X2.

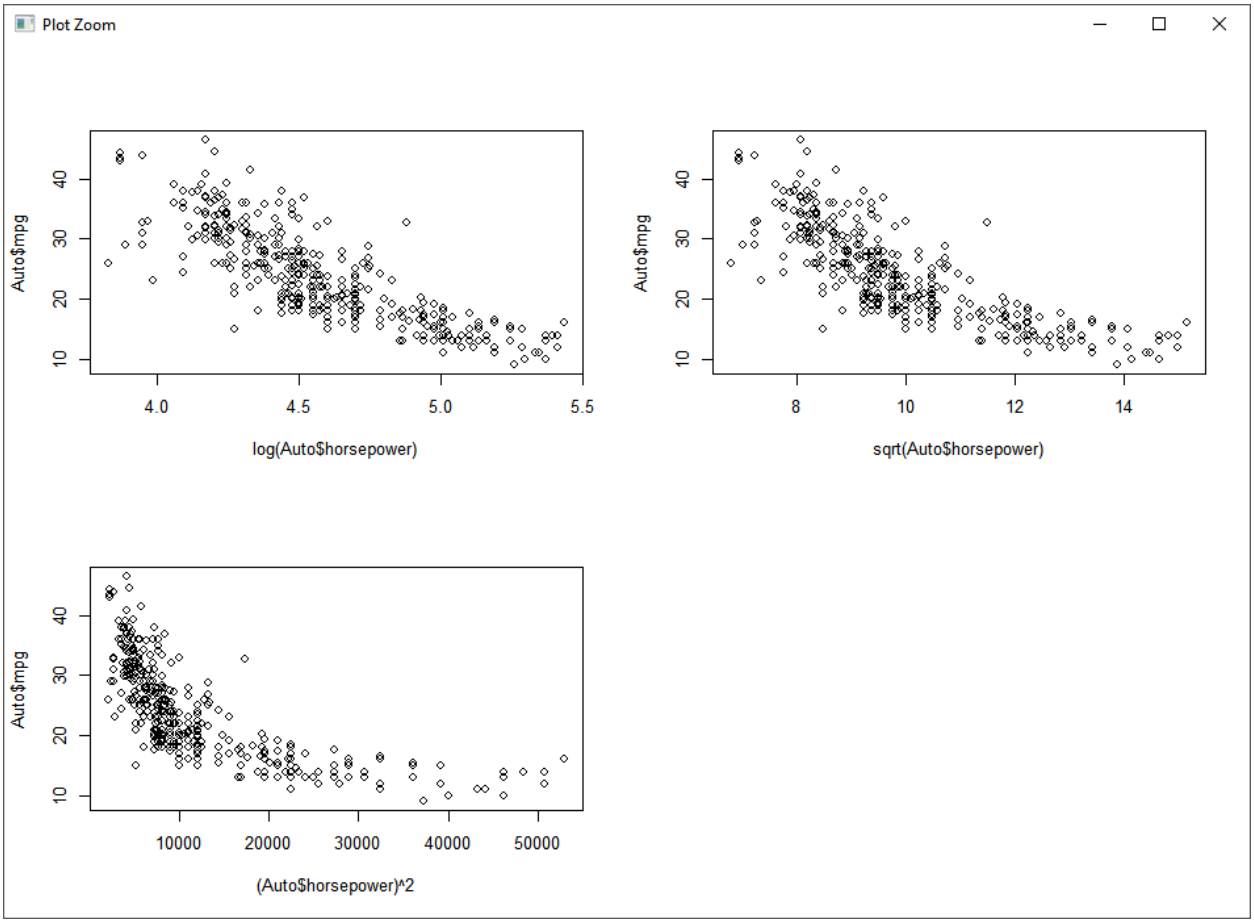
Comment on your findings

>par(mfrow = c(2, 2))

>plot(log(Auto$horsepower), Auto$mpg)

>plot(sqrt(Auto$horsepower), Auto$mpg)

>plot((Auto$horsepower)^2, Auto$mpg)



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

4. This question should be answered using the "Carseats" data set.

(a) Fit a multiple regression model to predict "Sales" using "Price", "Urban" and "US".

Ans.

Here the carseats is a simulated data set containing sales of child car seats at 400 different stores

We use Sales are determined by using price, Urban and US fields of the data frame.

>data(Carseats)

>fit3 <- lm(Sales ~ Price + Urban + US, data = Carseats)

>summary(fit3)

Where the Sales = (Price + Urban + US) + C

```
> 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
-6.9206 -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
```

(b) Provide an interpretation of each coefficient in the model. Be carefulsome 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 54.4588492 units in sales all other predictors remaining fixed. The coefficient of the "Urban" variable may be interpreted by saying that on average the unit sales in urban location are 21.9161508 units less than in rural location all other predictors remaining fixed. The coefficient of the "US" variable may be interpreted by saying that on average the unit sales in a US store are 1200.5726978 units more than in a non US store all other predictors remaining fixed.

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

The model may be written as

$$Sales=13.0434689+(-0.0544588)\times Price+(-0.0219162)\times Urban+(1.2005727)\times US+\varepsilon$$

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 $H0:\beta j=0$?

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)

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

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 ?

The R2 for the smaller model is marginally better than for the bigger model. Essentially about 23.9262888% of the variability is explained by the model.

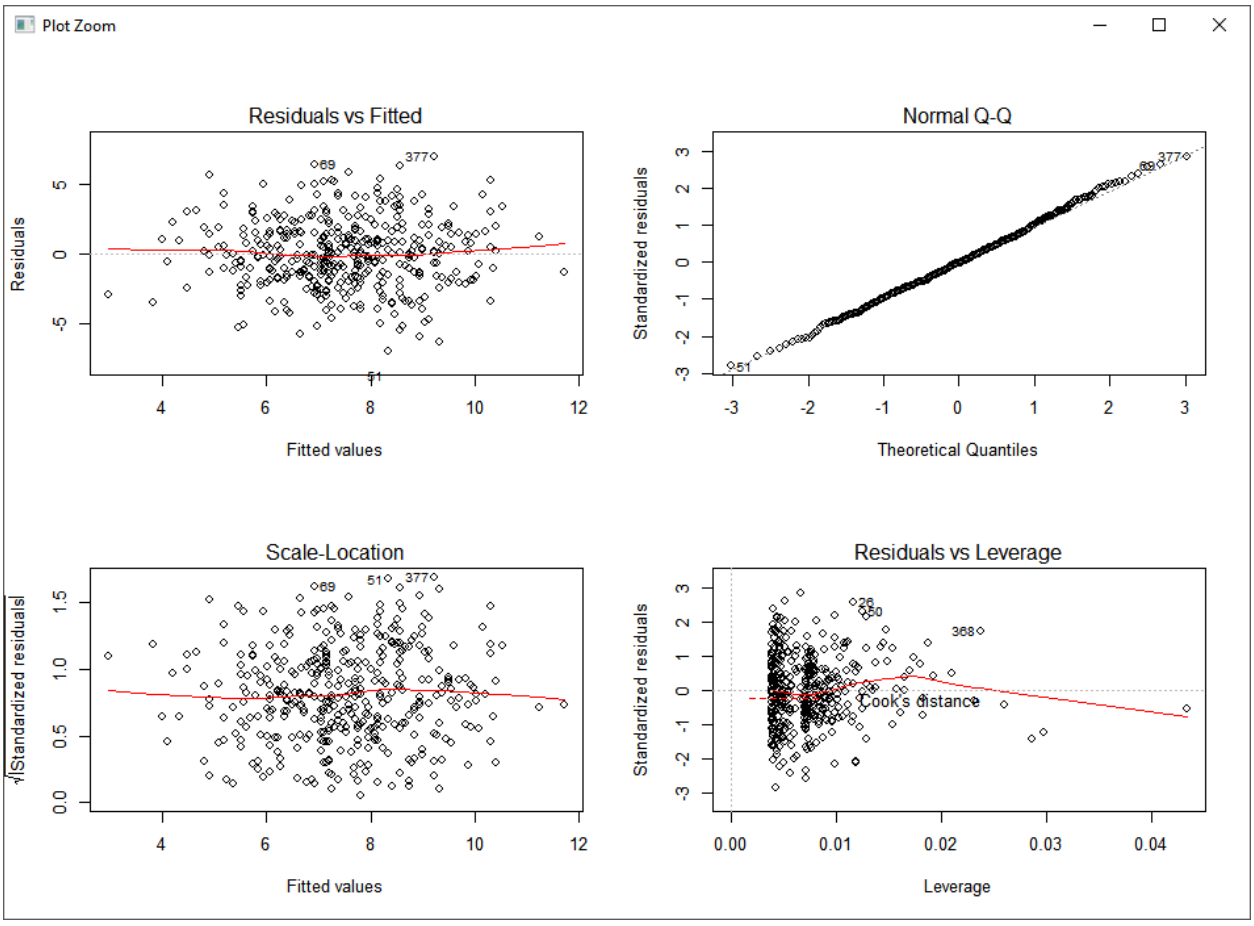
(g)Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
> confint(fit4)
                2.5 %      97.5 %
(Intercept) 11.79032020 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) ?

>par(mfrow = c(2, 2))

>plot(fit4)



The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and some leverage points as some points exceed (p+1)/n (0.01).

5. In this problem we will investigate the t-statistic for the null hypothesis H0:β=0 in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

(a) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate β^, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis H0. Comment on these results.

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

Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9154 -0.6472 -0.1771  0.5056  2.3109

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x   1.9939     0.1065   18.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

According to the summary above, we have a value of 1.9938761 for β^, a value of 0.1064767 for the standard error, a value of 18.7259319 for the t-statistic and a value of 2.642196910^{-34} for the p-value. The small p-value allows us to reject H0.

(b) Now perform a simple linear regression of x onto y, without an intercept. Report the coefficient estimate β^, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis H0. Comment on these results.

>fit6 <- lm(x ~ y + 0)

>summary(fit6)

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

Call:
lm(formula = x ~ y + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2368  0.1030  0.2858  0.8938

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y  0.39111    0.02089   18.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

According to the summary above, we have a value of 0.3911145 for β^β^, a value of 0.0208863 for the standard error, a value of 18.7259319 for the t-statistic and a value of 2.642196910^{-34} for the p-value. The small p-value allows us to reject H0H0.

(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. Both results in (a) and (b) reflect the same line created in (a). In other words, y=2x+εy=2x+ε could also be written x=0.5(y−ε)x=0.5(y−ε).

(d) For the regrssion of YY onto XX without an intercept, the t-statistic for H0:β=0H0:β=0 takes the form β^/SE(β^)β^/SE(β^), where β^β^ is given by (3.38), and where

```
> n <- length(x)
> t <- sqrt(n - 1)*(x %*% y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
> as.numeric(t)
[1] 18.72593
```

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 yy onto xx is the same t-statistic for the regression of xx onto yy.

It is easy to see that if we replace xixi by yiyi 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 yy onto xx as it is the regression of xx onto yy.

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

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8768 -0.6138 -0.1395  0.5394  2.3462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03769    0.09699  -0.389    0.698
x            1.99894    0.10773  18.556   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16

> fit8 <- lm(x ~ y)
> summary(fit8)

Call:
lm(formula = x ~ y)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90848 -0.28101  0.06274  0.24570  0.85736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03880    0.04266    0.91    0.365
y            0.38942    0.02099   18.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4249 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

It is again easy to see that the t-statistic for "fit7" and "fit8" are both equal to 18.5555993.