Abhishek HW2

Question 1

```
# (a) Perform a simple linear regression with mpg as the response and
horsepower as the predictor. Comment on the output.

library(ISLR)
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
```

```
fit = lm(mpg ~ horsepower, 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
```

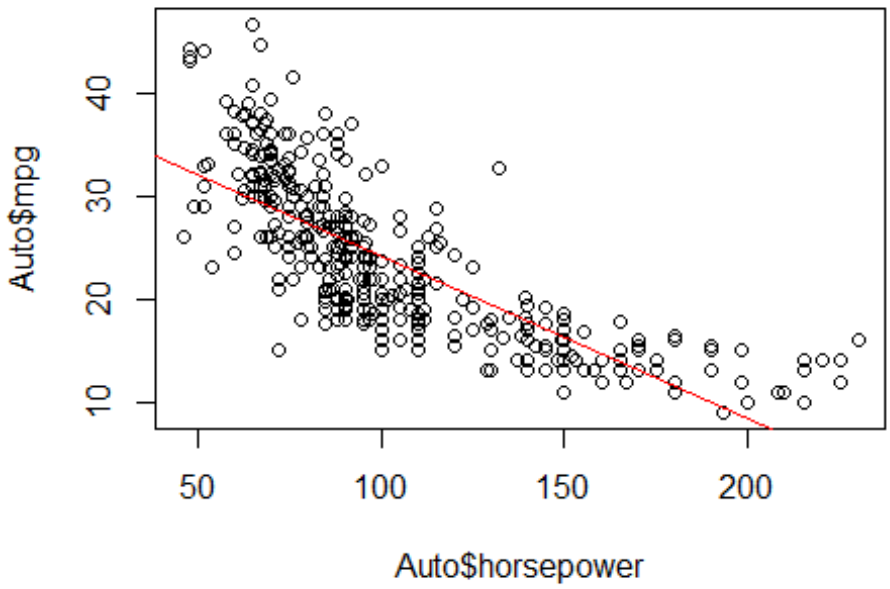
```r
# Is there a relationship between the predictor and the response?
# Yes. The p-value corresponding to the F-statistic is very low, indicating a
clear evidence of a relationship between mpg and horsepower.

# How strong is the relationship between the predictor and the response?
# Strong evidence of relationship, R2 statistic shows the percentage of
variability in the response that is explained by the predictors. The
predictors explain almost 60% of the variance in mpg.

# Is the relationship between the predictor and the response positive or
negative?
# Negative, since the coefficient has a negative value.

# How to interpret the estimate of the slope?
# If the horsepower increases by 1 unit, then mpg decreases by 0.16 unit.

# 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
```

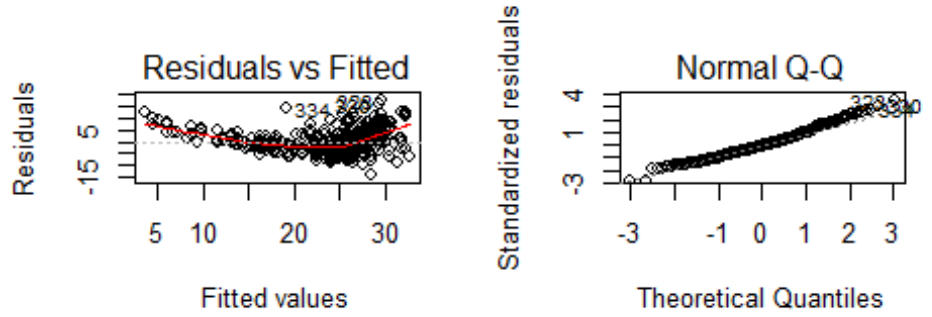
```r
predict(fit, data.frame(horsepower = 98), interval = "prediction")
```

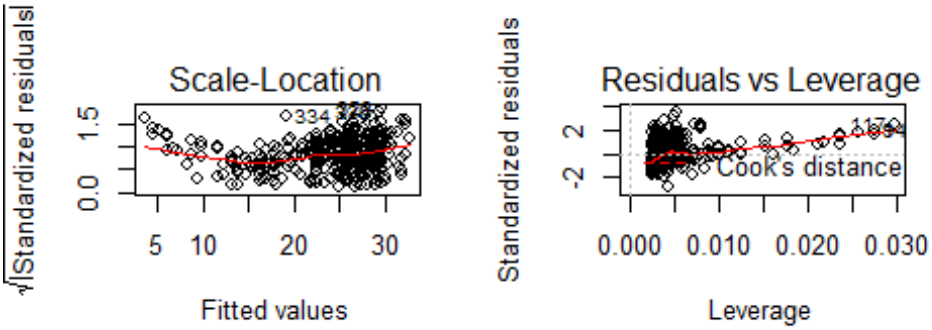
```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

```r
# 95% confidence interval is [23.97,24.96]
# 95% prediction interval is [14.8,34.12]


# (b) Plot the response and the predictor. Display the least squares
regression line in the plot.

plot(Auto$horsepower, Auto$mpg)
abline(fit, col = "red")
```




```r
# (c) Produce the diagnostic plots of the least squares regression fit.
Comment on each plot.
par(mfrow=c(2,2))
plot(fit, which=1)
plot(fit, which=2)
plot(fit, which=3)
plot(fit, which=5)
```






```
# The Residuals vs Fitted graph has a U-shape, thus the relationship between
predictors and response is nonlinear.
# The Residuals vs Fitted graph, it does not show heteroscedasticity.
# The Scale-Location graph indicates that there are outliers.
# The Residuals vs Leverage graph shows that there are many high leverage
points.

#log transformation
log_horsepower = log(Auto$horsepower)
log_fit = lm(mpg ~ log_horsepower, Auto)
plot(log_horsepower, Auto$mpg)
abline(log_fit, col = "red")
summary(log_fit)

##
## Call:
## lm(formula = mpg ~ log_horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2299  -2.7818  -0.2322   2.6661  15.4695
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    108.6997     3.0496   35.64   <2e-16 ***
## log_horsepower -18.5822     0.6629  -28.03   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.501 on 390 degrees of freedom
## Multiple R-squared:  0.6683, Adjusted R-squared:  0.6675
## F-statistic: 785.9 on 1 and 390 DF,  p-value: < 2.2e-16
```

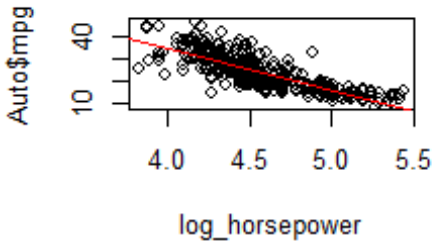
```
predict(log_fit, data.frame(log_horsepower = 98), interval = "confidence")
```

```
##         fit       lwr       upr
## 1 -1712.354 -1834.091 -1590.618
```

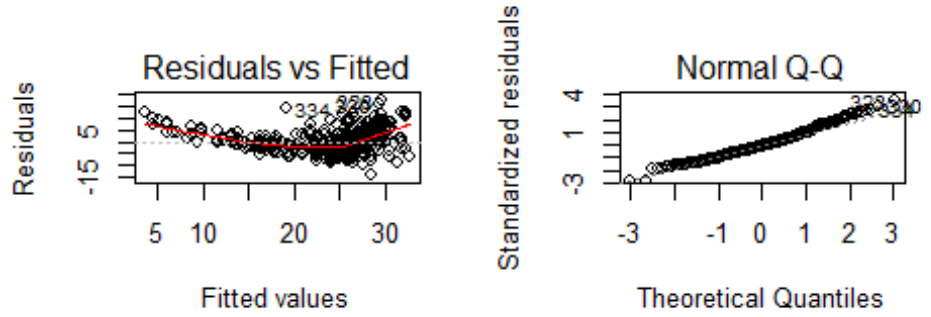
```
predict(log_fit, data.frame(log_horsepower = 98), interval = "prediction")
```

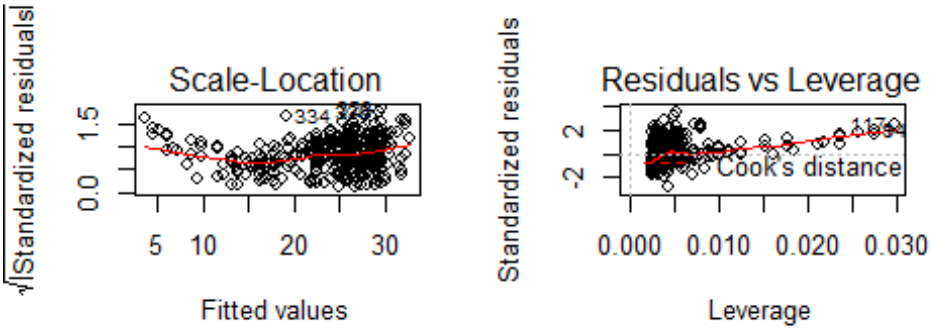
```
##         fit       lwr       upr
## 1 -1712.354 -1834.412 -1590.297
```

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




```
plot(fit, which=1)
plot(fit, which=2)
plot(fit, which=3)
plot(fit, which=5)
```






```
# R2 statistic is 66.8% and hence is a better fit compared to the model
without transformation.

#Square-root transformation
sqrt_horsepower = sqrt(Auto$horsepower)
sqrt_fit = lm(mpg ~ sqrt_horsepower, Auto)
plot(sqrt_horsepower, Auto$mpg)
abline(sqrt_fit, col = "red")
summary(sqrt_fit)

##
## Call:
## lm(formula = mpg ~ sqrt_horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9768  -3.2239  -0.2252   2.6881  16.1411
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       58.705      1.349   43.52   <2e-16 ***
## sqrt_horsepower   -3.503      0.132  -26.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.665 on 390 degrees of freedom
```

```
## Multiple R-squared:  0.6437, Adjusted R-squared:  0.6428
## F-statistic: 704.6 on 1 and 390 DF,  p-value: < 2.2e-16
```

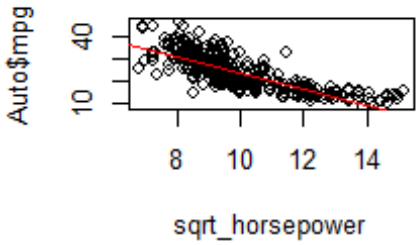
```
predict(sqrt_fit, data.frame(sqrt_horsepower = 98), interval = "confidence")
```

```
##         fit       lwr       upr
## 1 -284.6402 -307.4641 -261.8163
```

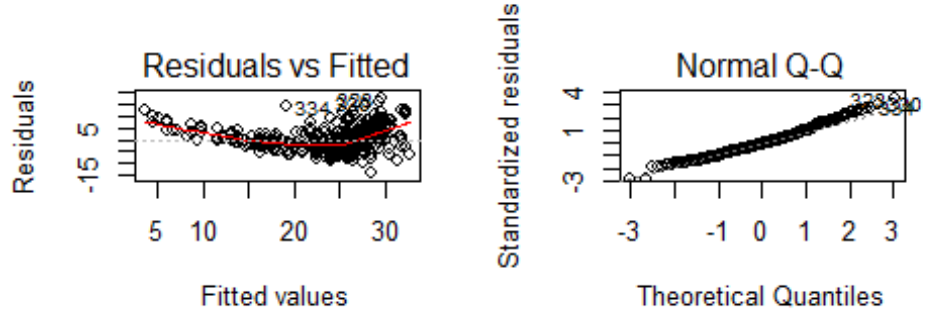
```
predict(sqrt_fit, data.frame(sqrt_horsepower = 98), interval = "prediction")
```

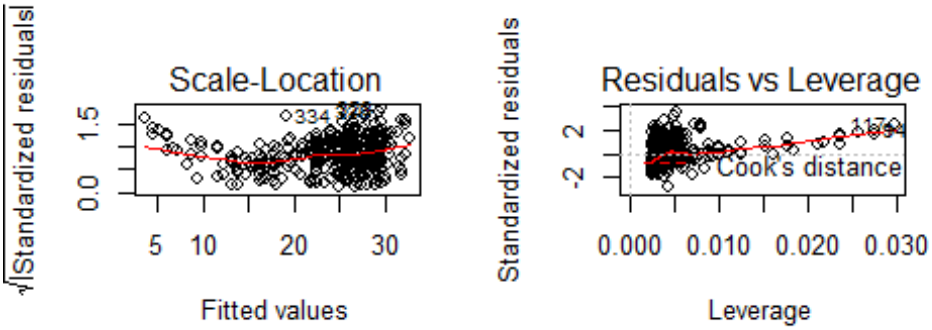
```
##         fit       lwr       upr
## 1 -284.6402 -309.2378 -260.0425
```

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




```
plot(fit, which=1)
plot(fit, which=2)
plot(fit, which=3)
plot(fit, which=5)
```






```
# R2 statistic is 64.3% and hence is a better fit compared to the model
without transformation.

#Square transformation
square_horsepower = (Auto$horsepower)^2
square_fit = lm(mpg ~ square_horsepower, Auto)
plot(square_horsepower, Auto$mpg)
abline(square_fit, col = "red")
summary(square_fit)

##
## Call:
## lm(formula = mpg ~ square_horsepower, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.529  -3.798  -1.049   3.240  18.528
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.047e+01  4.466e-01   68.22   <2e-16 ***
## square_horsepower -5.665e-04  2.827e-05  -20.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.485 on 390 degrees of freedom
```

```
## Multiple R-squared:  0.5074, Adjusted R-squared:  0.5061
## F-statistic: 401.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

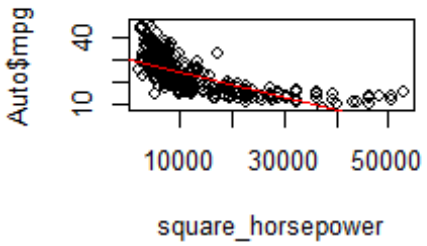
```
predict(square_fit, data.frame(square_horsepower = 98), interval =
"confidence")
```

```
##        fit     lwr      upr
## 1 30.41026 29.5365 31.28401
```

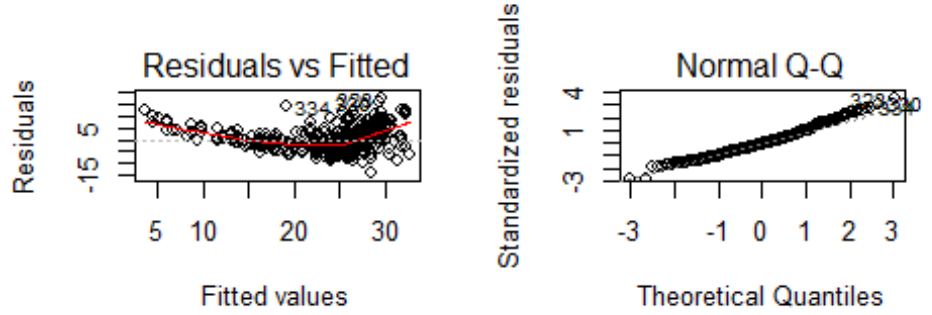
```
predict(square_fit, data.frame(square_horsepower = 98), interval =
"prediction")
```

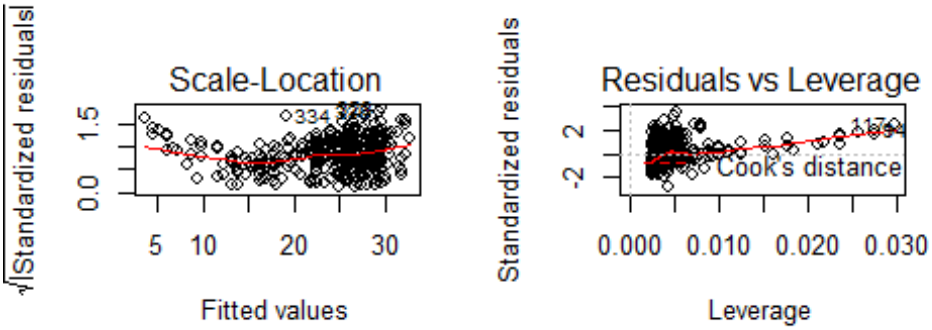
```
##        fit      lwr      upr
## 1 30.41026 19.59069 41.22982
```

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




```
plot(fit, which=1)
plot(fit, which=2)
plot(fit, which=3)
plot(fit, which=5)
```





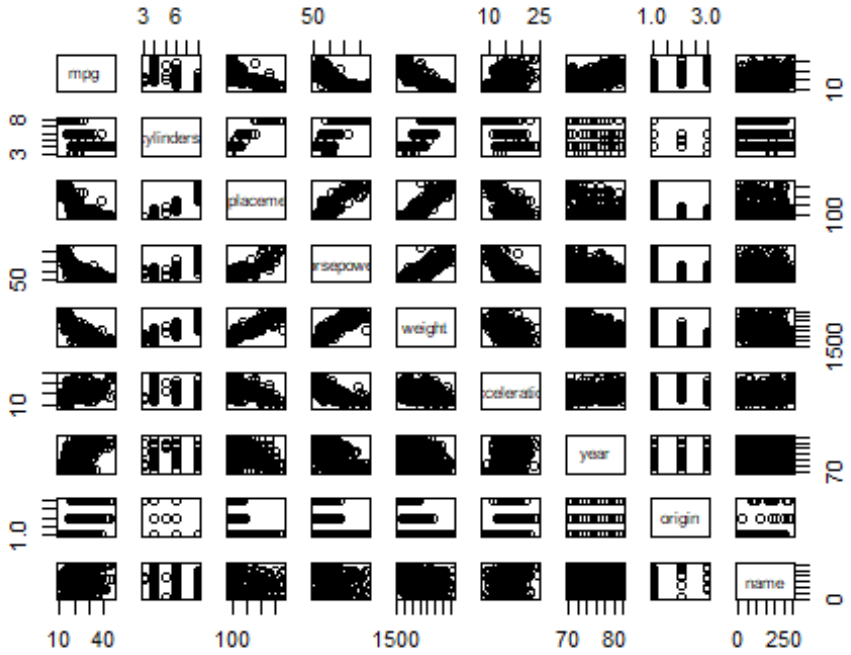

```
# R2 statistic is 50.7% and hence is not a better fit compared to the model
without transformation.
```

Question 2

```
# (a) Produce a scatterplot matrix which includes all of the variables in the
data set. Which predictors appear to have an association with the response?
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
```

```
pairs(Auto)
```




```
# mpg vs horsepower, mpg vs weight, dispacement vs weight, weight vs
horsepower, weight vs mpg are correlated.

# (b) Compute the matrix of correlations between the variables (using the
function cor()). You will need to exclude the name variable, which is
qualitative.
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
```

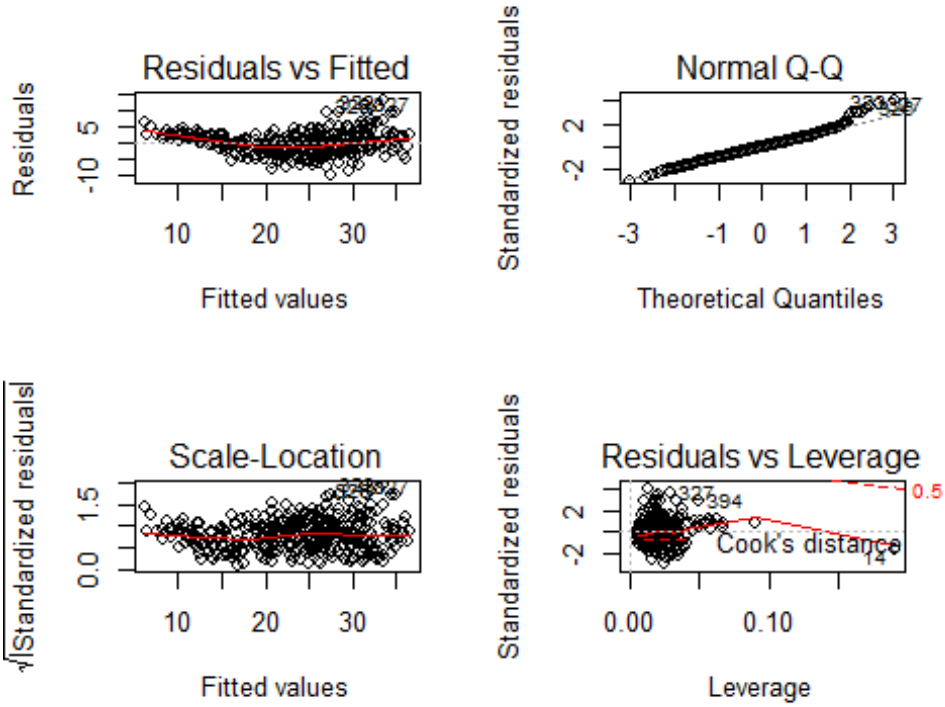
```
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

```
# (c) Perform a multiple linear regression with mpg as the response and all
other variables except name as the predictors. Comment on the output. For
example,
fit = lm(mpg ~ .-name, data=Auto)
summary(fit)
```

```
##
## 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) Is there a relationship between the predictors and the response?
# Yes, A large F-statistic and the corresponding small p-value indicates that
there is a relationship between .

# ii) Which predictors appear to have a statiscally significant relationship
to the response?
# Displacement, Weight, Year and Origin.

# iii) What does the coefficient for the year variable suggest?
# For each additional year, more 0.75 miles per galon is possible.

# (d) Produce diagnostic plots of the linear regression fit. Comment on each
plot.
par(mfrow=c(2,2))
plot(fit, which=1)
plot(fit, which=2)
```

```r
plot(fit, which=3)
plot(fit, which=5)
```




```r
# The Residuals vs Fitted graph does not have a U-shape curve hence the possibility of non-linear relationship can be eliminated .
# The Residuals vs Fitted graph takes a funnel shape indicates non-constant variance of errors.
# The Scale-Location graph shows that there are outliers.
# The Residuals vc Leverage graph showa that observation 14 is a high leverage point.

# (e) Is there serious collinearity problem in the model? Which predictors are collinear?
library(car)

## Loading required package: carData

vif(fit)

##    cylinders displacement   horsepower       weight acceleration
##    10.737535    21.836792     9.943693    10.831260     2.625806
##         year       origin
##     1.244952     1.772386

# A value of VIF>5 indicates serious collinearity. The predictors cylinders, displacement, horsepower and weight contrubute to collinearity problem.
# COllinearity reduces the accuracy of the estimates of the regressioN coefficients.
```

```r
# (f) Fit linear regression models with interactions. Are any interactions
statistically significant?
fit_inter = lm(mpg ~ (.-name)*(.-name), data=Auto)
summary(fit_inter)
```

```
##
## Call:
## lm(formula = mpg ~ (. - name) * (. - name), data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.548e+01  5.314e+01   0.668  0.50475
## cylinders                    6.989e+00  8.248e+00   0.847  0.39738
## displacement                -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                   5.034e-01  3.470e-01   1.451  0.14769
## weight                       4.133e-03  1.759e-02   0.235  0.81442
## acceleration                -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                         6.974e-01  6.097e-01   1.144  0.25340
## origin                      -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders:displacement      -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower         1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight             3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration       2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year              -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin             4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower     -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight          2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration   -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year            5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin          2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight           -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration     -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year             -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin            2.233e-03  2.930e-02   0.076  0.93931
## weight:acceleration          2.346e-04  2.289e-04   1.025  0.30596
## weight:year                 -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin               -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year            5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin          4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                  1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
```

```
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16

# The p value for acceleration:origin interaction is less than 0.05 and hence
this interaction is statistically significant.
```

Question 3

```
head(Carseats)

##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes

fit_1 = lm(Sales ~ Price + Urban + US, data=Carseats)
summary(fit_1)

##
## 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
```

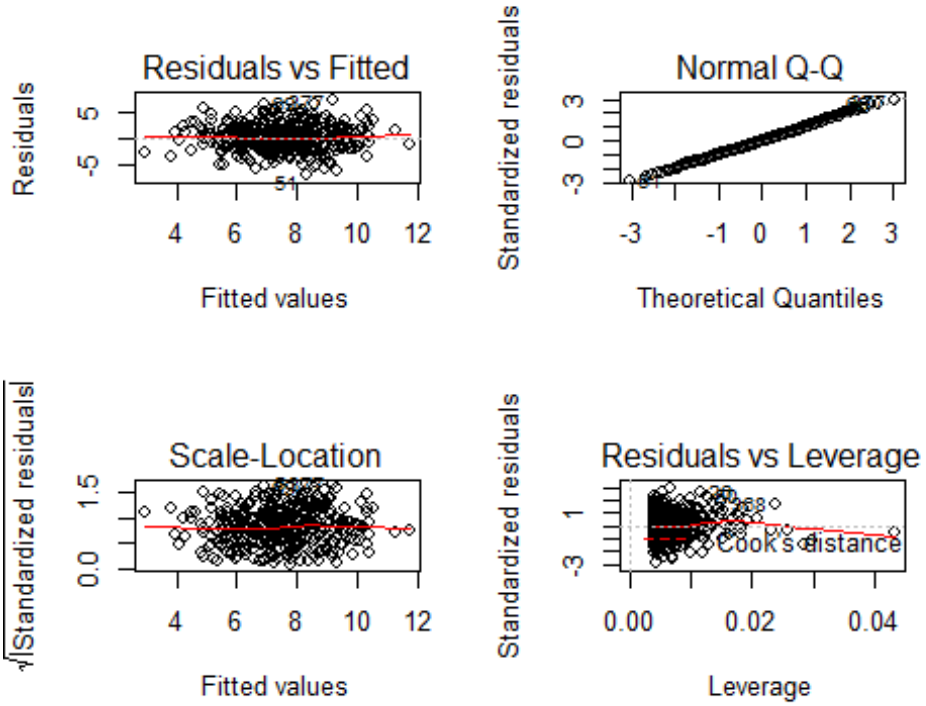
```
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16

# (b) Provide an interpretation of each coefficient in the model
# The UrbanYes has a very high p-value hence this predictor can be neglected.
# The USYes has a very low p-value hence this predictor cannot be negelected.
An additional 1.2 thousands sales units is assigned for a US location.
# The Price has a negative relationship with Sales

# (c) Write out the model in equation form.
# Sales = 13.043-0.055*Price-0.022*UrbanYes+1.2*USYes

# (d) For which of the predictors can you reject the null hypothesis 𝐻0: 𝛽𝑗=0
?
# We can reject the null hypothesis for Price & US predictors .

# (e) On the basis of your answer to the previous question, fit a smaller
model that only uses the predictors for which there is evidence of
association with the response.
fit_2 = lm(Sales ~ Price + US, data=Carseats)
summary(fit_2)

##
## 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?
summary(fit_1)

##
## 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
```

```
summary(fit_2)
```

```
##
## 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
```

```
# The R2 statistic is same for both fits. The F-statistic is large for the
second fit and hence is a superior fit.

# (g) Is there evidence of outliers or high leverage observations in the
model from (e)?
par(mfrow=c(2,2))
plot(fit_2, which=1)
plot(fit_2, which=2)
plot(fit_2, which=3)
plot(fit_2, which=5)
```




```
# Scale-Location graph does not show any highlighted outlier.
# Residuals vs Leverage graph shows a very high leverage observation.
```