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.

```
library("ISLR")

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

i. Is there a relationship between the predictor and the response?
The p-values for the regression coefficients are nearly zero. This implies statistical significance, which in turn mean that there is a relationship.

ii. How strong is the relationship between the predictor and the response?
The R^{2} value indicates that about 61% of the variation in the response variable (mpg) is due to the predictor variable (horsepower).

iii. Is the relationship between the predictor and the response positive or negative? The regression coefficient for 'horsepower' is negative. Hence, the relationship is negative.

iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?
The confidence 95% interval

```
predict(lm.fit, data.frame(horsepower = c(85)), interval ="confidence")
##        fit    lwr      upr
## 1 26.51906 25.973 27.06512
```

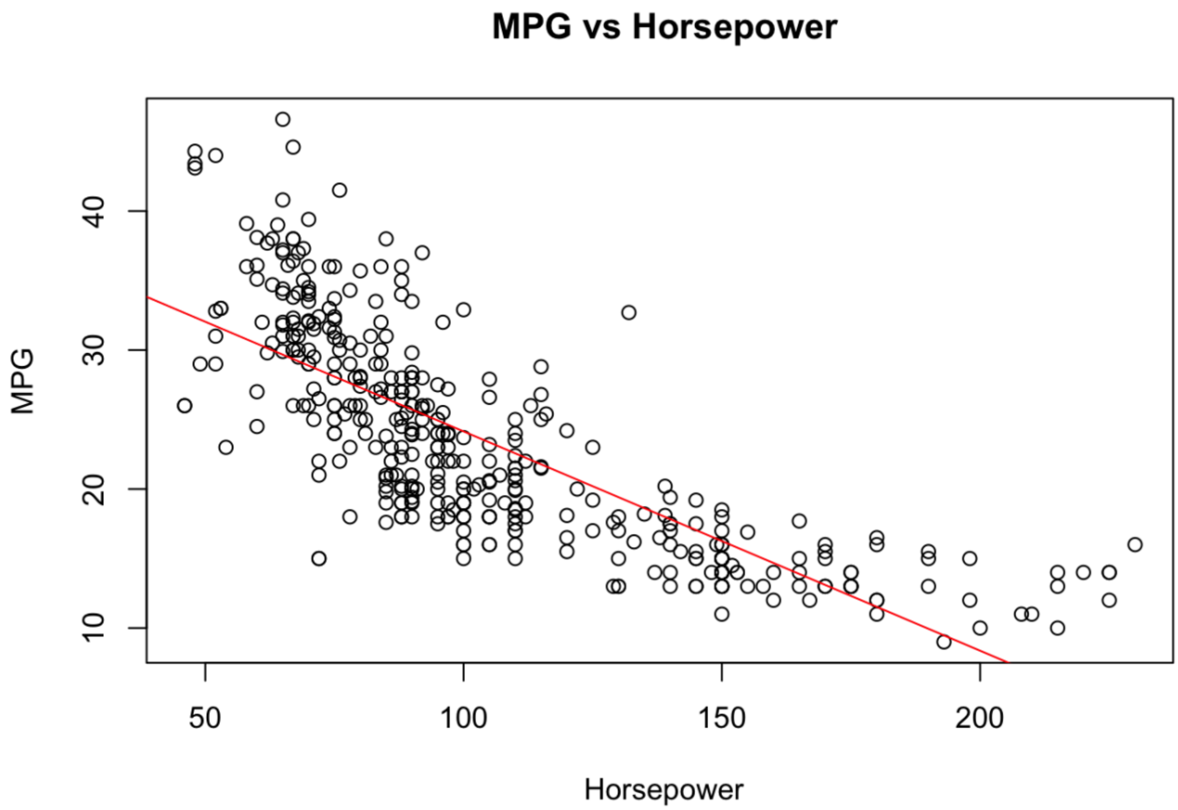
And, the 95% prediction interval

```
predict(lm.fit, data.frame(horsepower = c(85)), interval ="prediction")
##        fit      lwr      upr
## 1 26.51906 16.85857 36.17954
```

As expected the prediction interval is wider than the confidence interval.
Plot the response and the predictor.

   b. Use the abline() function to display the least squares regression line.

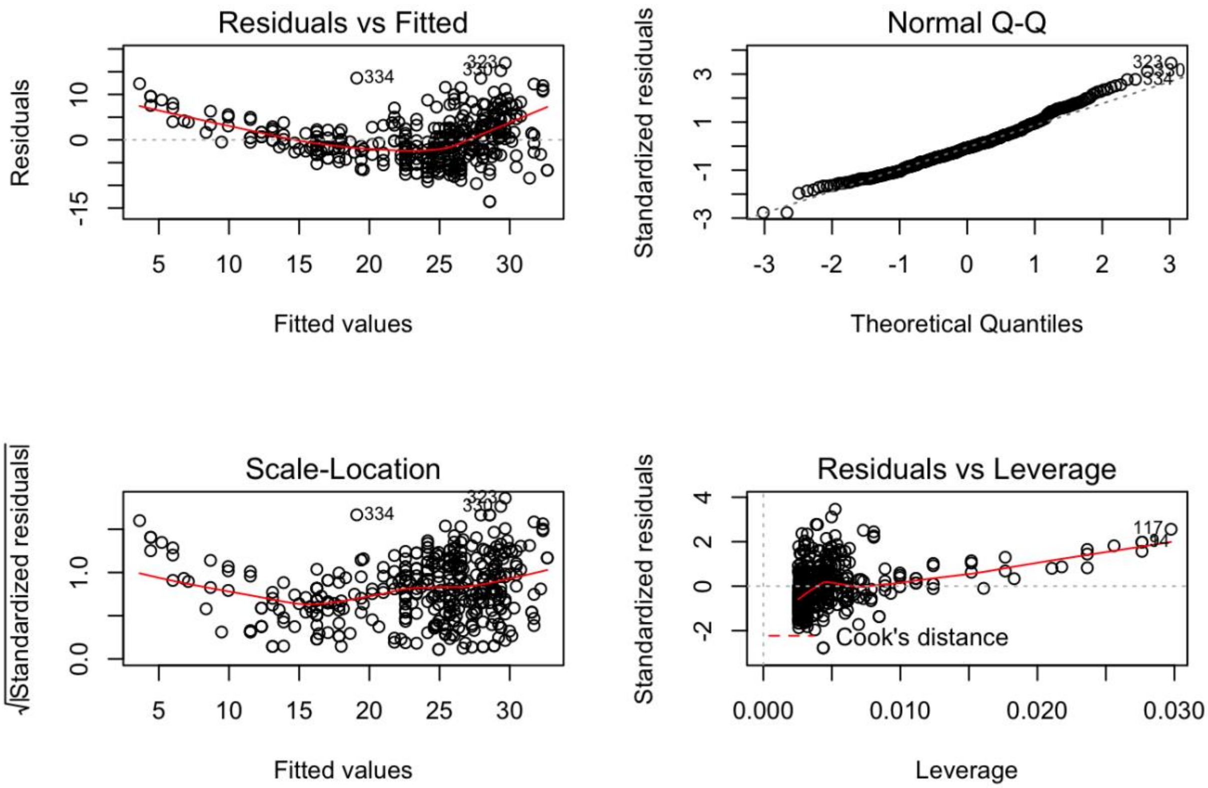
```
attach(Auto)
plot(mpg~horsepower, main =" MPG vs Horsepower", xlab = " Horsepower", ylab
="MPG")
abline(coef = coef(lm.fit), col ="red")
```




```
detach(Auto)
```

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(lm.fit)
```

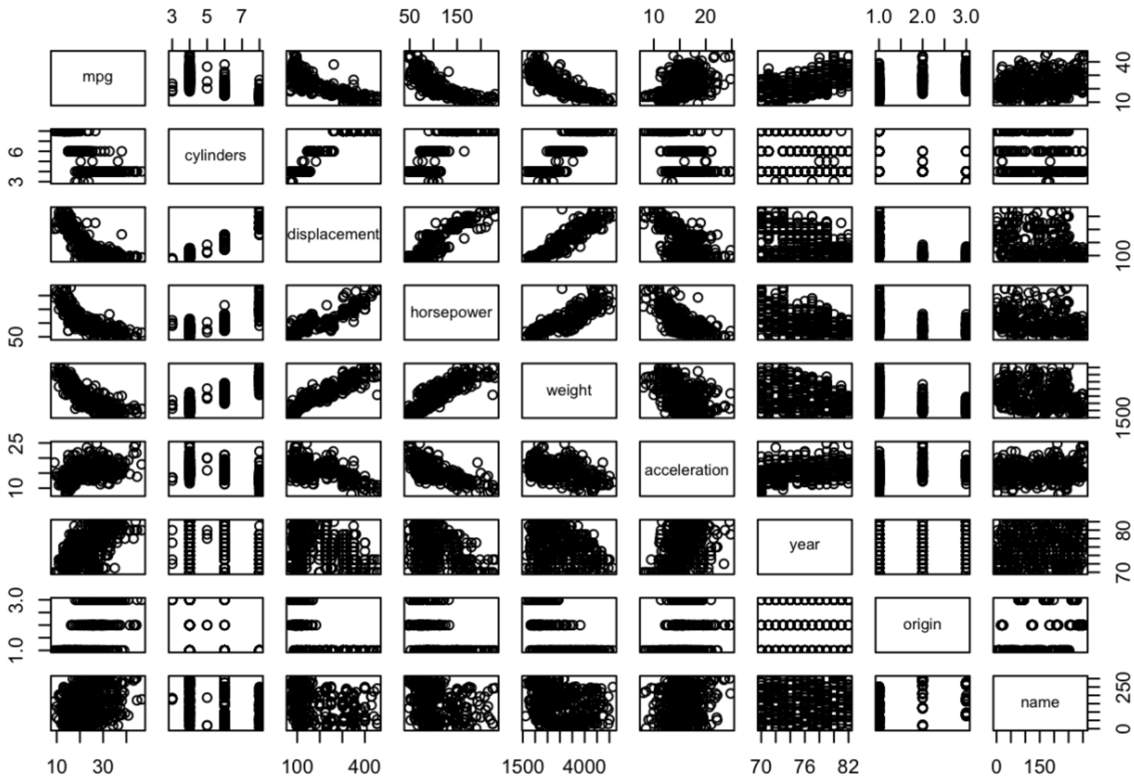


The first plot shows a pattern (U-shaped) between the residuals and the fitted values. This indicates a non-linear relationship between the predictor and response variables. The second plot shows that the residuals are normally distributed. The third plot shows that the variance of the errors is constant. Finally, the fourth plot indicates that there are no leverage points in the data.

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

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

```
library("ISLR")
pairs(Auto)
```



a. 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[, names(Auto) !="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
```

b. 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:

```
model = lm(mpg ~. -name, data = Auto)
summary(model)
##
## 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
```

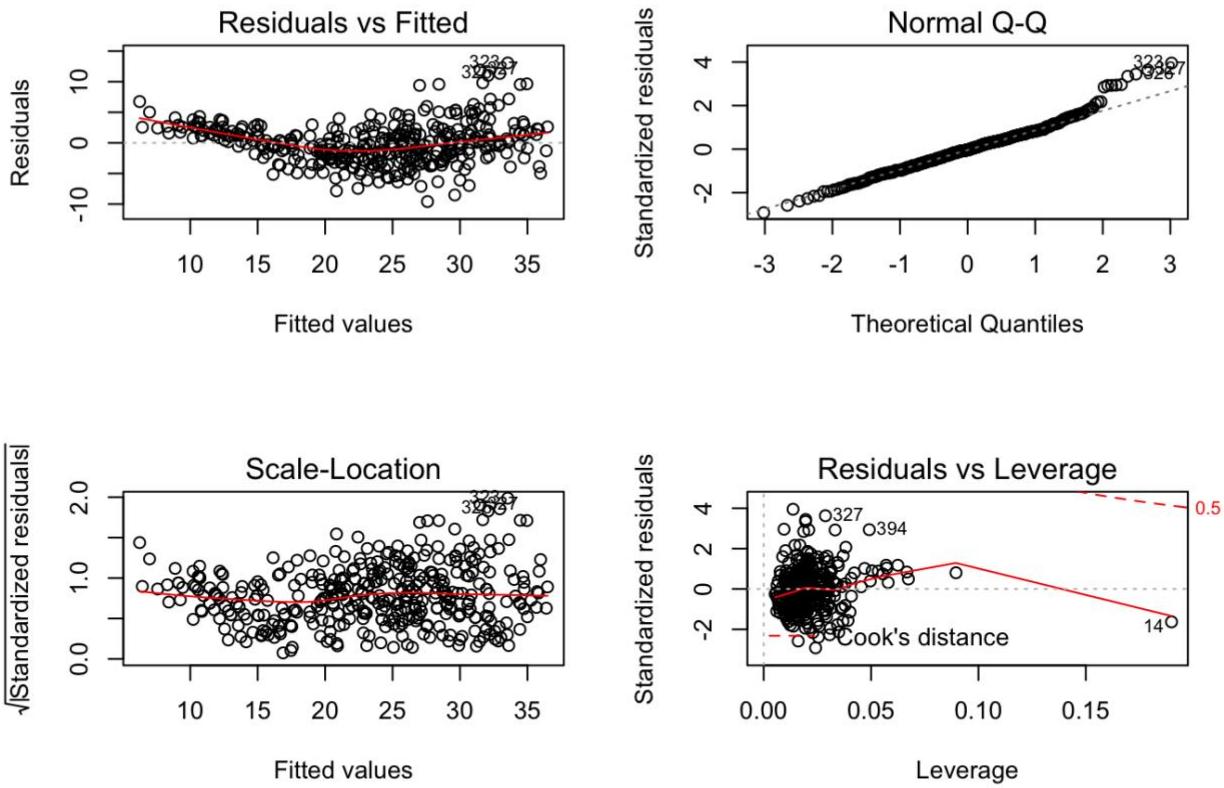
i. Is there a relationship between the predictors and the response? Yes, there is. However, some predictors do not have a statistically significant effect on the response. R-squared value implies that 82% of the changes in the response can be explained by the predictors in this regression model.

ii. Which predictors appear to have a statistically significant relationship to the response? displacement, weight, year, origin .

iii. What does the coefficient for the year variable suggest? When every other predictor held constant, the mpg value increases with each year that passes. Specifically, mpg increase by 1.43 each year.

c. 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(model)
```



The first graph shows that there is a non-linear relationship between the responce and the predictors;
The second graph shows that the residuals are normally distributed and right skewed;
The third graph shows that the constant variance of error assumption is not true for this model;
The Third graphs shows that there are no leverage points. However, there on observation that stands out as a potential leverage point (labeled 14 on the graph)

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

```
model = lm(mpg ~.-name+displacement:weight, data = Auto)
summary(model)
##
## Call:
## lm(formula = mpg ~ . - name + displacement:weight, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9027 -1.8092 -0.0946  1.5549 12.1687
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -5.389e+00  4.301e+00  -1.253   0.2109
## cylinders            1.175e-01  2.943e-01   0.399   0.6899
## displacement        -6.837e-02  1.104e-02  -6.193 1.52e-09 ***
## horsepower          -3.280e-02  1.238e-02  -2.649   0.0084 **
```

```
## weight                -1.064e-02  7.136e-04 -14.915  < 2e-16 ***
## acceleration           6.724e-02  8.805e-02   0.764   0.4455
## year                   7.852e-01  4.553e-02  17.246  < 2e-16 ***
## origin                 5.610e-01  2.622e-01   2.139   0.0331 *
## displacement:weight    2.269e-05  2.257e-06  10.054  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.964 on 383 degrees of freedom
## Multiple R-squared:  0.8588, Adjusted R-squared:  0.8558
## F-statistic: 291.1 on 8 and 383 DF,  p-value: < 2.2e-16
```

e. model = lm(mpg ~.-name+displacement:cylinders+displacement:weight+acceleration:horsepower, data=Auto)

```
summary(model)
##
## Call:
## lm(formula = mpg ~ . - name + displacement:cylinders + displacement:weight
+
##     acceleration:horsepower, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3344 -1.6333  0.0188  1.4740 11.9723
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.725e+01  5.328e+00  -3.237  0.00131 **
## cylinders               6.354e-01  6.106e-01   1.041  0.29870
## displacement           -6.805e-02  1.337e-02  -5.088 5.68e-07 ***
## horsepower              6.026e-02  2.601e-02   2.317  0.02105 *
## weight                 -8.864e-03  1.097e-03  -8.084 8.43e-15 ***
## acceleration            6.257e-01  1.592e-01   3.931  0.00010 ***
## year                    7.845e-01  4.470e-02  17.549  < 2e-16 ***
## origin                  4.668e-01  2.595e-01   1.799  0.07284 .
## cylinders:displacement -1.337e-03  2.726e-03  -0.490  0.62415
## displacement:weight     2.071e-05  3.638e-06   5.694 2.49e-08 ***
## horsepower:acceleration -7.467e-03  1.784e-03  -4.185 3.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.905 on 381 degrees of freedom
## Multiple R-squared:  0.865,  Adjusted R-squared:  0.8615
## F-statistic: 244.2 on 10 and 381 DF,  p-value: < 2.2e-16
```

f. model = lm(mpg ~.-name+displacement:cylinders+displacement:weight+year:origin+acceleration:horsepower, data=Auto)

```
summary(model)
##
## Call:
```

```
## lm(formula = mpg ~ . - name + displacement:cylinders + displacement:weight
+
##     year:origin + acceleration:horsepower, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6504 -1.6476  0.0381  1.4254 12.7893
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.287e+00  9.074e+00   0.583 0.560429
## cylinders                 4.249e-01  6.079e-01   0.699 0.485011
## displacement             -7.322e-02  1.334e-02  -5.490 7.38e-08 ***
## horsepower                5.252e-02  2.586e-02   2.031 0.042913 *
## weight                   -8.689e-03  1.086e-03  -7.998 1.54e-14 ***
## acceleration              5.796e-01  1.582e-01   3.665 0.000283 ***
## year                      5.116e-01  9.976e-02   5.129 4.66e-07 ***
## origin                   -1.220e+01  4.161e+00  -2.933 0.003560 **
## cylinders:displacement   -4.368e-04  2.712e-03  -0.161 0.872156
## displacement:weight       1.992e-05  3.608e-06   5.522 6.21e-08 ***
## year:origin               1.630e-01  5.341e-02   3.051 0.002440 **
## horsepower:acceleration -6.735e-03  1.781e-03  -3.781 0.000181 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.874 on 380 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8644
## F-statistic: 227.7 on 11 and 380 DF,  p-value: < 2.2e-16
```

g. model = lm(mpg ~.-name-cylinders-acceleration+year:origin+displacement:weight+ displacement:weight+acceleration:horsepower+acceleration:weight, data=Auto)
summary(model)

```
##
## Call:
## lm(formula = mpg ~ . - name - cylinders - acceleration + year:origin +
##     displacement:weight + displacement:weight + acceleration:horsepower +
##     acceleration:weight, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5074 -1.6324  0.0599  1.4577 12.7376
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.868e+01  7.796e+00   2.396 0.017051 *
## displacement             -7.794e-02  9.026e-03  -8.636  < 2e-16 ***
## horsepower                8.719e-02  3.167e-02   2.753 0.006183 **
## weight                   -1.350e-02  1.287e-03 -10.490  < 2e-16 ***
## year                      4.911e-01  9.825e-02   4.998 8.83e-07 ***
## origin                   -1.262e+01  4.109e+00  -3.071 0.002288 **
## year:origin               1.686e-01  5.277e-02   3.195 0.001516 **
## displacement:weight       2.253e-05  2.184e-06  10.312  < 2e-16 ***
```

```
## horsepower:acceleration -9.164e-03  2.222e-03  -4.125 4.56e-05 ***
## weight:acceleration       2.784e-04  7.087e-05   3.929 0.000101 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.861 on 382 degrees of freedom
## Multiple R-squared:  0.8687, Adjusted R-squared:  0.8656
## F-statistic: 280.8 on 9 and 382 DF,  p-value: < 2.2e-16
```

From all the 4 models, the last model is the only one with all variables being significant. And, based on results from a few trials not show here, it is very likely that it is the best combination of predictors and interaction terms. The R-squared statistics estimates that 87% of the changes in the response can be explained by this particular set of predictors (single and interaction.) A higher value was not obtained from the trials.

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.*

```
library("ISLR")
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

str(Carseats)
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2
3 3 ...
##  $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...

lm.fit = lm(Sales ~ Price+Urban+US, data= Carseats)
```

```
summary(lm.fit)
##
## 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 careful—some of the variables in the model are qualitative!*

When price increases by $1000 and other predictors are held constant, sales decrease by 54.459 unit sales. In otherwords, when price increases by $1000, the number of carseats sold decrease by 54,459.
A store's sale is not affected by whether or not it is in a Urban area.
A store in the US sales 1200 more carseats (in average) than a store that is abroad.

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

Skipped.

d. *For which of the predictors can you reject the null hypothesis Ho : βj = 0?*

The predictor 'Urban'. Its p-value is not statistically significant with a value of 0.936.

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.

```
lm.fit2 = lm(Sales ~ Price+US, data= Carseats)
summary(lm.fit2)
##
## 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?

Based on their respective R-square values(in summary tables), these two models are mediocre (only 24% change in response explained).

g.

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

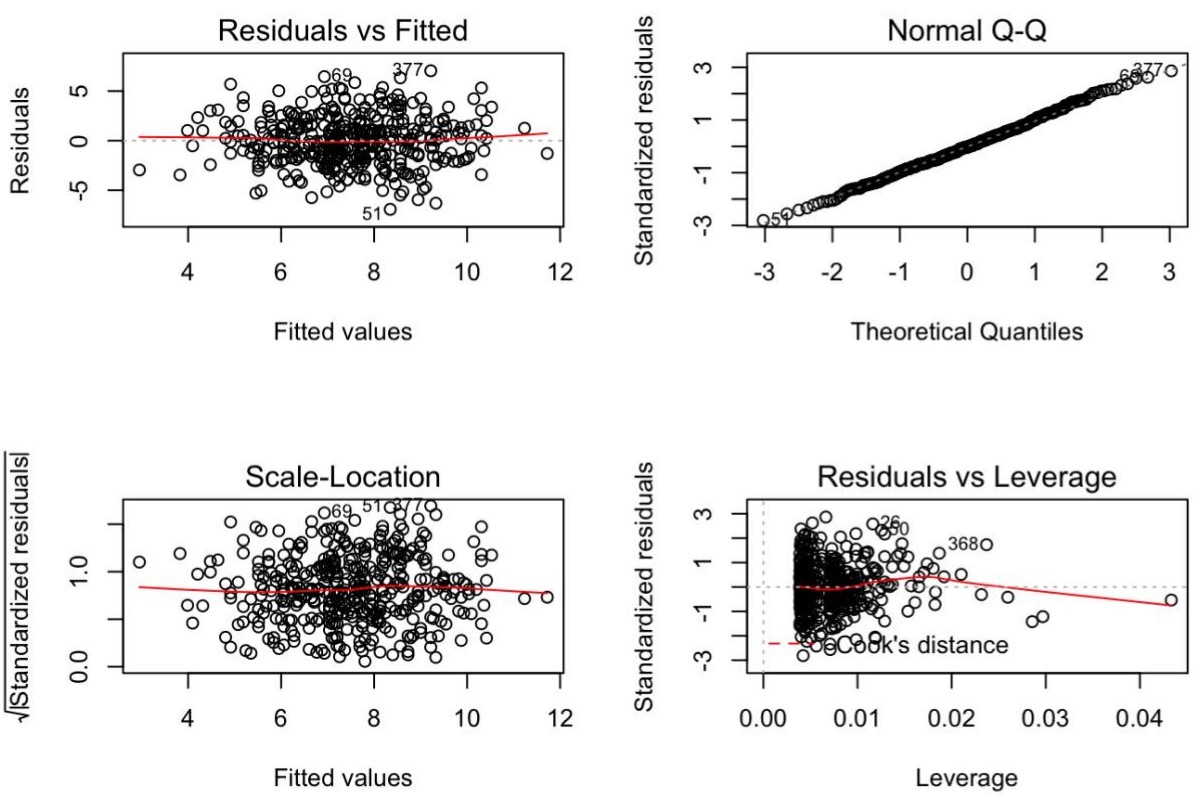
```
confint(lm.fit2)
```

```
##                   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(lm.fit2)
```



Based on the Normal.q-q pot and the Residuals vs Leverage plot, there are no evidence of such points.

5.In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
> set.seed(1)
> x = rnorm(100)
>y=2*x+rnorm (100)
```

a)

```
set.seed(1)
x=rnorm(100)
y=2*x+rnorm(100)
slr<-lm(y~x+0)
summary(slr)
##
## 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
```

coefficient estimate is **1.9938761** and standard error is 0.1065.The t statistic is obtained by dividing paramter estimate with its standard error which is given by 18.73 and p-value associated with is less than 0.05 and hence we reject the null hypothesis that coefficient of x is zero.

b)

```
revslr<-lm(x~y+0)
summary(revslr)
##
## 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
```

coefficient estimate is **0.3911145** and standard error is 0.02089.The t statistic is obtained by dividing paramter estimate with its standard error which is given by 18.73 and p-value associated with is less than 0.05 and hence we reject the null hypothesis that coefficient of x is zero.

c)

We get the same t statistic and p-value for both the cases and the intercept is changed and it is not the inverse so we cannot say that y=mx+c is written as x=(1/m)(y-c)

d)

```
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 get t-statistic as **18.7259319** from the formula which is equal to the t-statistic that we obtained earlier by dividing parameter estimate beta by standard error of beta.

e)

As the formula indicates it only depends on the value of x and y we get same t statistic for both the cases

f)

```
revslr1<-lm(x~y)
summary(revslr1)
##
## 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
```

```
slr1<-lm(y~x)
summary(slr1)
```

```
##
## 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
```

We get the t-statistic as **18.56** for both the cases.