PREDICT 422 Assignment 2

Artur Mrozowski

April 2, 2017

PREDICT 422 Programming Assignment 2

Excercise 9 (ISLR Section 10.7)

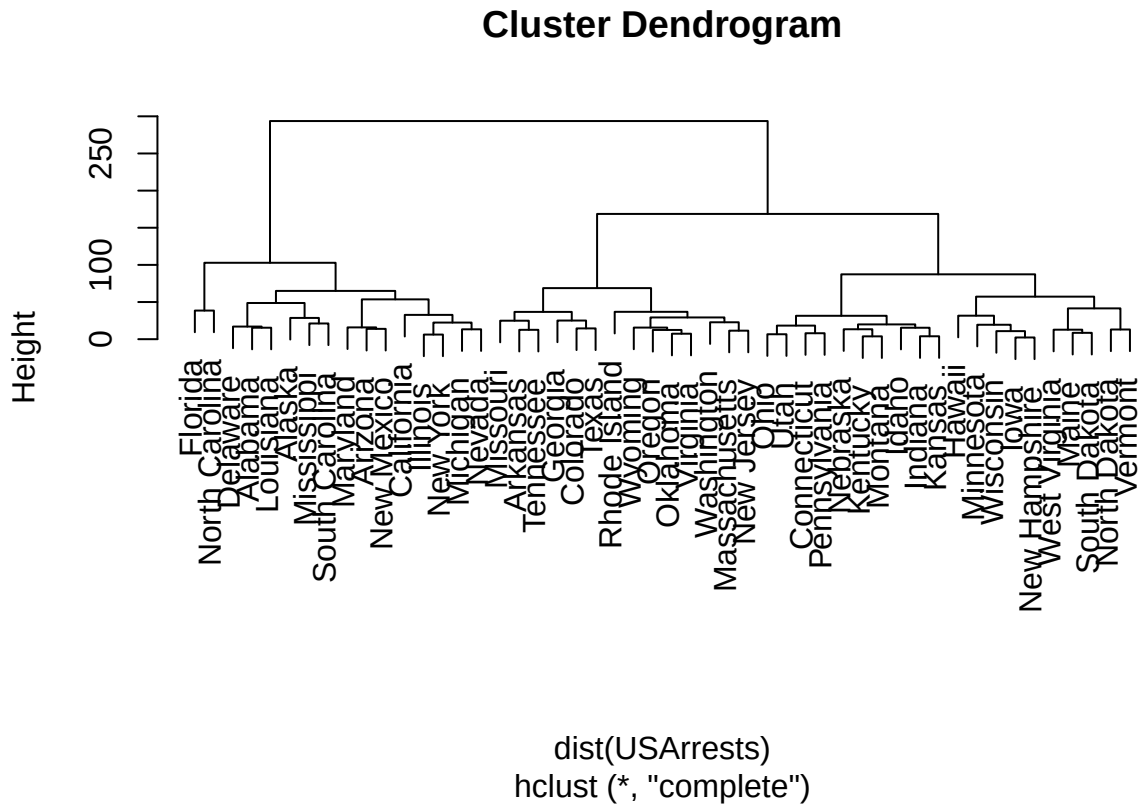
```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.2.3
```

```
set.seed(2)
```

9(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

```
hc.complete = hclust(dist(USArrests), method="complete")
plot(hc.complete)
```



9(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```r
cutree(hc.complete, 3)
```

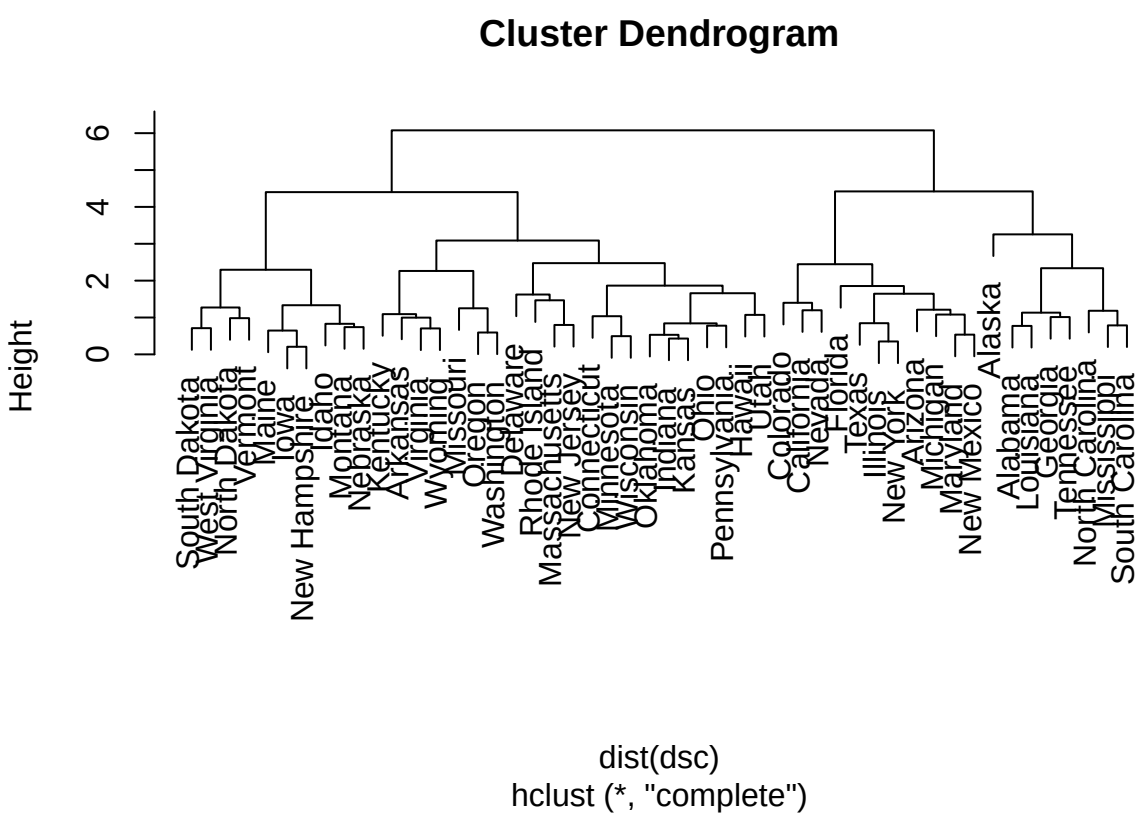
```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

```r
table(cutree(hc.complete, 3))
```

```
##
##  1  2  3
## 16 14 20
```

9(c)Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

```r
dsc = scale(USArrests)
hc.s.complete = hclust(dist(dsc), method="complete")
plot(hc.s.complete)
```



9(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

```
cutree(hc.s.complete, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

```
table(cutree(hc.s.complete, 3))
```

```
##
##  1  2  3
##  8 11 31
```

```
table(cutree(hc.s.complete, 3), cutree(hc.complete, 3))
```

```
##
##      1  2  3
##   1  6  2  0
##   2  9  2  0
##   3  1 10 20
```

Scaling the variables effects the max height of the dendogram obtained from hierarchical clustering. From a cursory glance, it doesn't effect the bushiness of the tree obtained. However, it does affect the clusters obtained from cutting the dendogram into 3 clusters. UrbanPop is variable describing percent of population in urban area and is not comparable to number of assault per 100000. Variables should be scaled to have standard mean and deviation.

10.

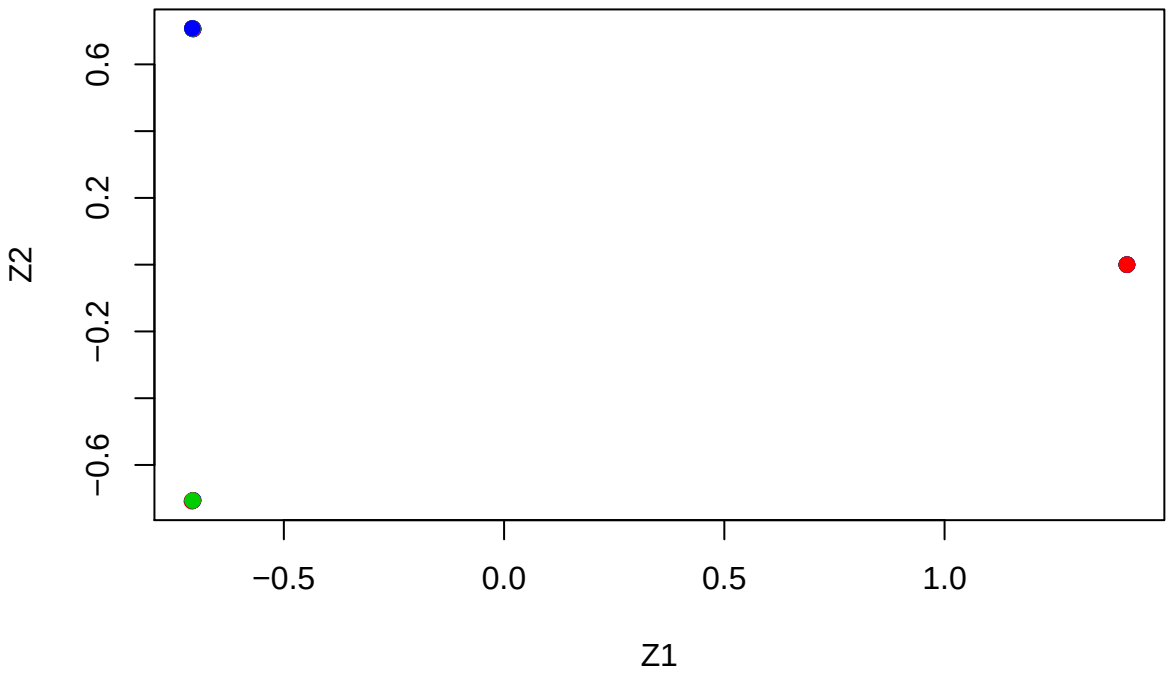
(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables. Hint: There are a number of functions in R that you can use to generate data. One example is the rnorm() function; runif() is another option. Be sure to add a mean shift to the observations in each class so that there are three distinct classes.

```
set.seed(2)
x = matrix(rnorm(20*3*50, mean=0, sd=0.001), ncol=50)
x[1:20, 2] = 1
x[21:40, 1] = 2
x[21:40, 2] = 2
x[41:60, 1] = 1
```

(b) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes.

```
pca.out = prcomp(x)
summary(pca.out)
pca.out$x[,1:2]
```

```
plot(pca.out$x[,1:2], col=2:4, xlab="Z1", ylab="Z2", pch=19)
```



(c) Perform K-means clustering of the observations with K = 3. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

```
km.out = kmeans(x, 3, nstart=20)
table(km.out$cluster, c(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1 20  0  0
##   2  0 20  0
##   3  0  0 20
```

(d) Perform K-means clustering with K = 2.

```
km.out = kmeans(x, 2, nstart=20)
km.out$cluster
```

One of the previous classes absored into the other.

```
table(km.out$cluster)
```

```
##
##  1  2
## 40 20
```

(e) Now perform K-means clustering with K = 4, and describe your results.

```
km.out = kmeans(x, 4, nstart=20)
km.out$cluster
```

```
##  [1] 4 4 4 4 4 1 4 1 4 4 1 1 1 1 4 1 4 1 1 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [36] 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

One of the clusters split into two.

```
table(km.out$cluster)
```

```
##
##  1  2  3  4
##  9 20 20 11
```

(f) Now perform K-means clustering with K = 3 on the first two principal component score vectors, rather than on the raw data.

```
km.out = kmeans(pca.out$x[,1:2], 3, nstart=20)
table(km.out$cluster, c(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1  0  0 20
##   2  0 20  0
##   3 20  0  0
```

Perfect match just like before

(g) Using the scale() function, perform K-means clustering with K = 3 on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in (b)? Explain.

```
km.out = kmeans(scale(x), 3, nstart=20)
km.out$cluster
```

```
##  [1] 1 1 1 1 1 3 3 3 2 3 1 3 3 3 1 3 2 3 3 1 2 2 2 2 2 2 2 2 2 2 2 3 2 2
## [36] 2 2 2 2 2 1 1 1 3 1 3 1 1 1 3 2 3 3 1 1 3 1 3 3 1
```

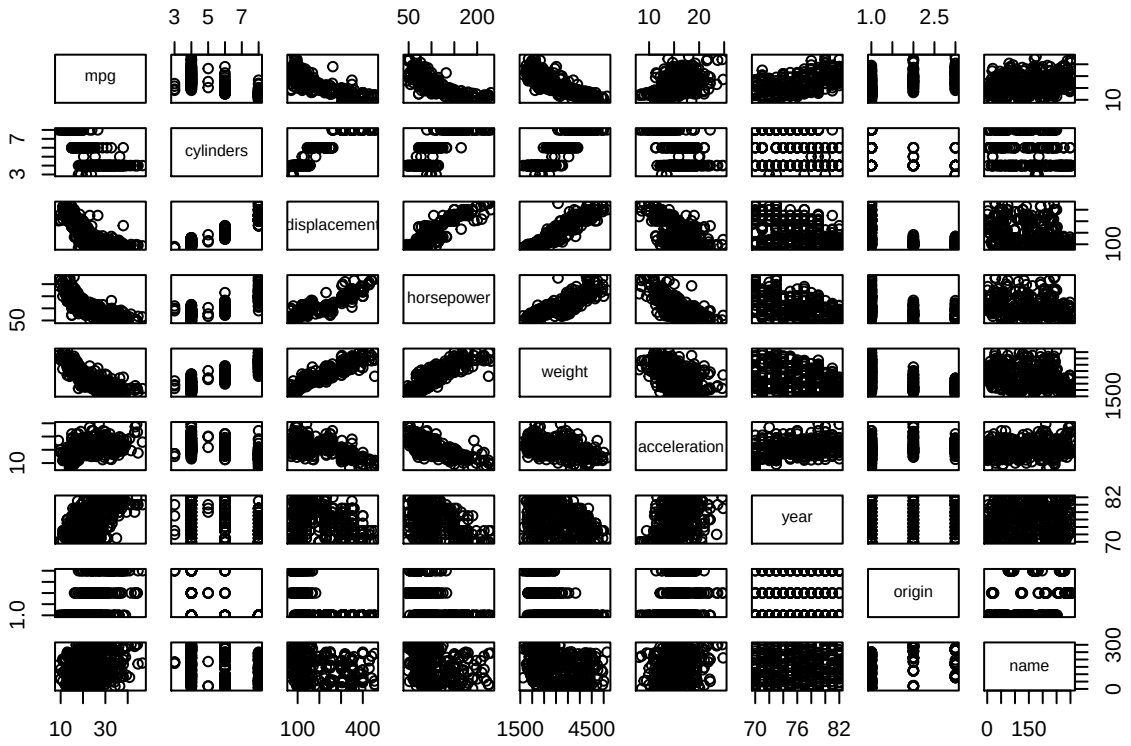
```
table(km.out$cluster, c(rep(1,20), rep(2,20), rep(3,20)))
```

```
##
##      1  2  3
##   1  8  0 11
##   2  2 19  1
##   3 10  1  8
```

The classification is not as distinct as in b. The scale function evenes out the distances between points.

Chapter 3 9 (a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
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.

```
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:

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?

```
lm.fit1 = lm(mpg~.-name, data=Auto)
summary(lm.fit1)
```

```
##
## 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. Yes there is a relationship between predictors and response. From F statistics we can reject the null hypothesis that the all predictor variable coefficients are equal to zero. F=252.4 and p-value 2.2e-16.

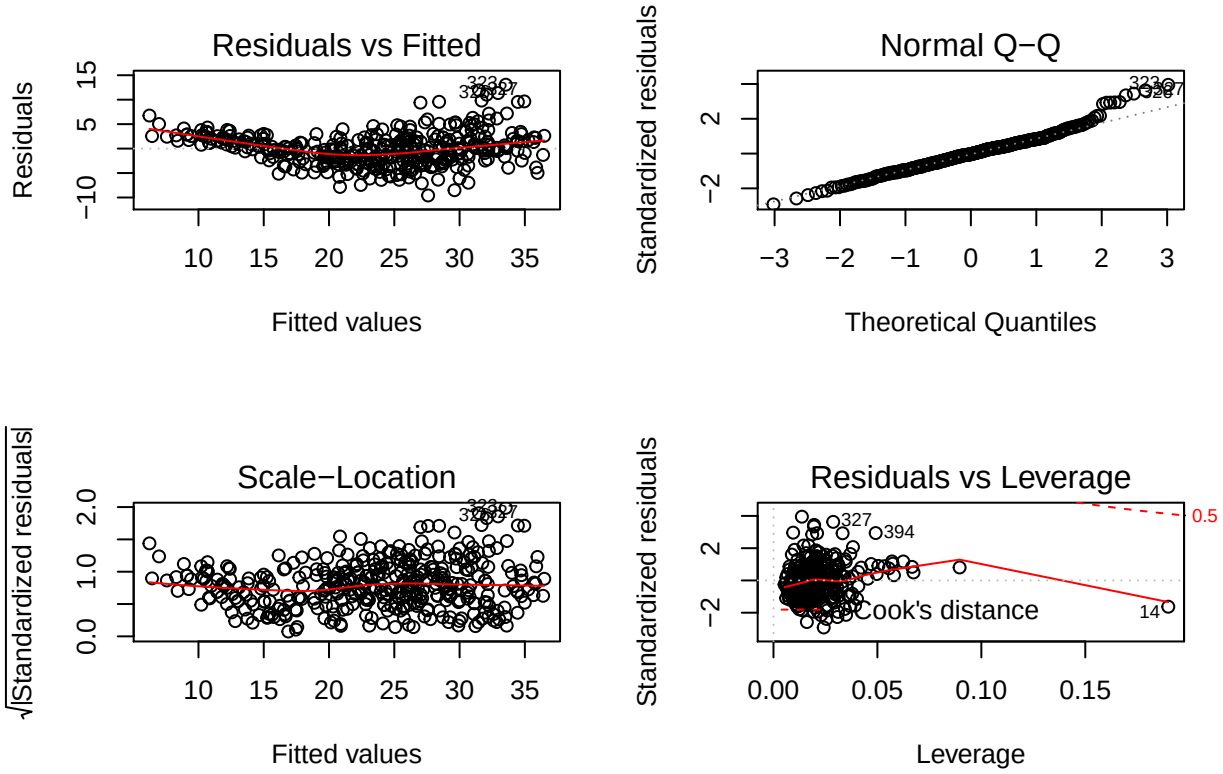
ii. Looking at the p-values associated with each predictor's t-statistic, we see that displacement, weight, year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not.

iii. The year coefficient 0.75 indicates that each year the cars become more and more efficient.

(d)

(e) 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(lm.fit1)
```



The analysis of Q-Q plot reveals that the right tail is a littble bit thick which indicates that there may be outliers. The analysis of residuals indicates that there is a curve pattern in the residual plot which may indicate non linear realtionship not explained by the model. That might require transformation of some of the variables.

From the leverage plot point 14 has high leverage which means that the observation is influential. There seems to be a few outliers higher than 2 and lower than -2

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

```
lm.fit2 = lm(mpg~cylinders*displacement+displacement*weight, data=Auto[,1:8])
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto[, 1:8])
##
## 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
```

From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interactiion between cylinders and displacement is not.

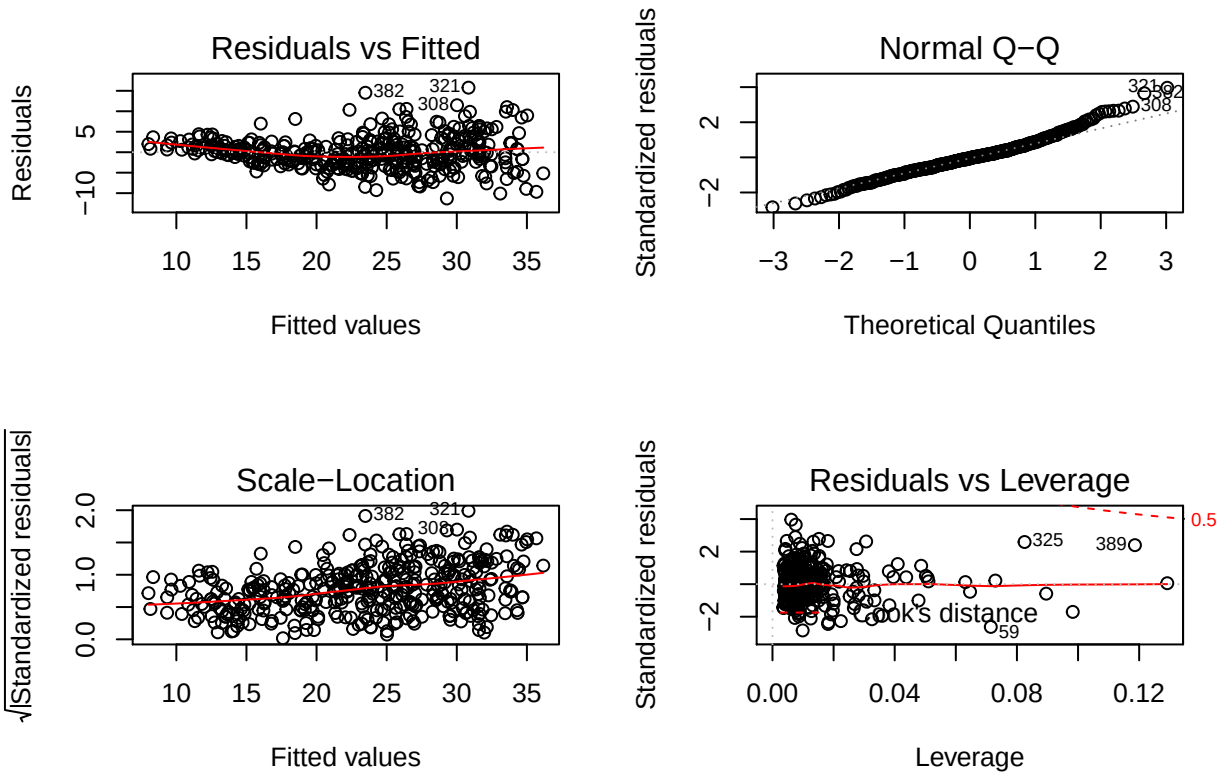
(f)

(g) Try a few different transformations of the variables, such as log(X),sqrt X, X2. Comment on your findings.

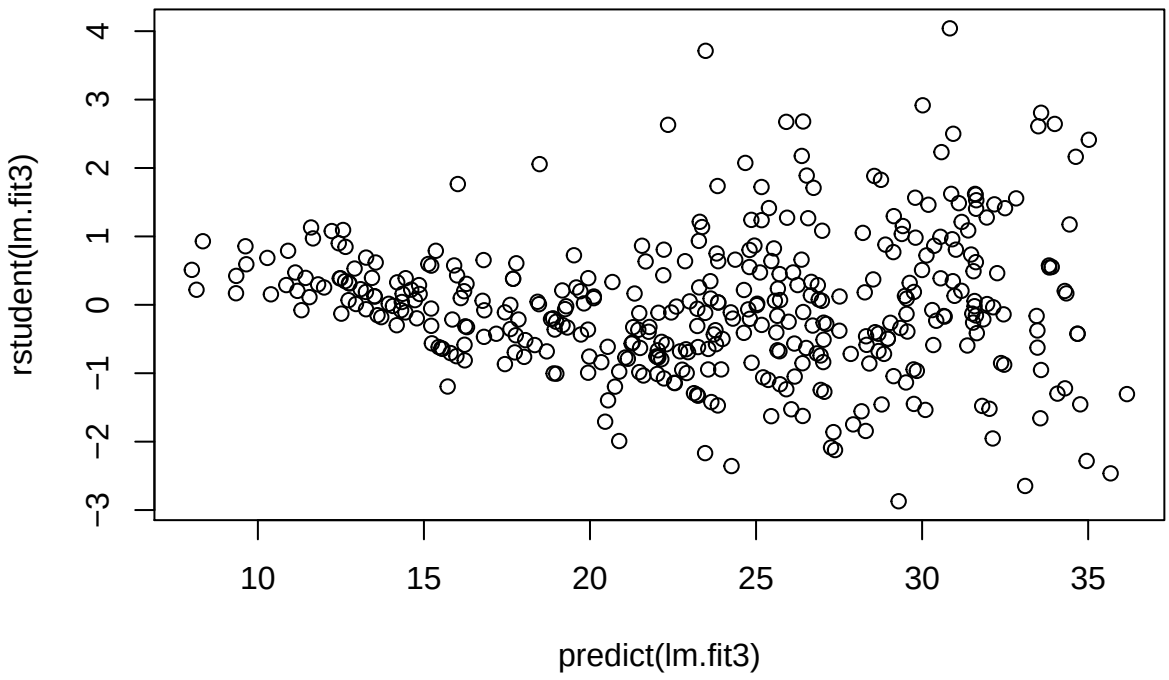
```
attach(Auto)
lm.fit3 = lm(mpg~log(weight)+sqrt(horsepower)+acceleration+I(acceleration^2))
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = mpg ~ log(weight) + sqrt(horsepower) + acceleration +
##     I(acceleration^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2932  -2.5082  -0.2237   2.0237  15.7650
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       178.30303   10.80451  16.503  < 2e-16 ***
## log(weight)       -14.74259    1.73994  -8.473 5.06e-16 ***
## sqrt(horsepower)   -1.85192    0.36005  -5.144 4.29e-07 ***
## acceleration       -2.19890    0.63903  -3.441 0.000643 ***
## I(acceleration^2)   0.06139    0.01857   3.305 0.001037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.99 on 387 degrees of freedom
## Multiple R-squared:  0.7414, Adjusted R-squared:  0.7387
## F-statistic: 277.3 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm.fit3)
```




```
plot(predict(lm.fit3), rstudent(lm.fit3))
```



2 problems are observed from the above plots: 1) the residuals vs fitted plot indicates heteroskedasticity (unconstant variance over mean) in the model. 2) The Q-Q plot indicates somewhat unnormality of the residuals.

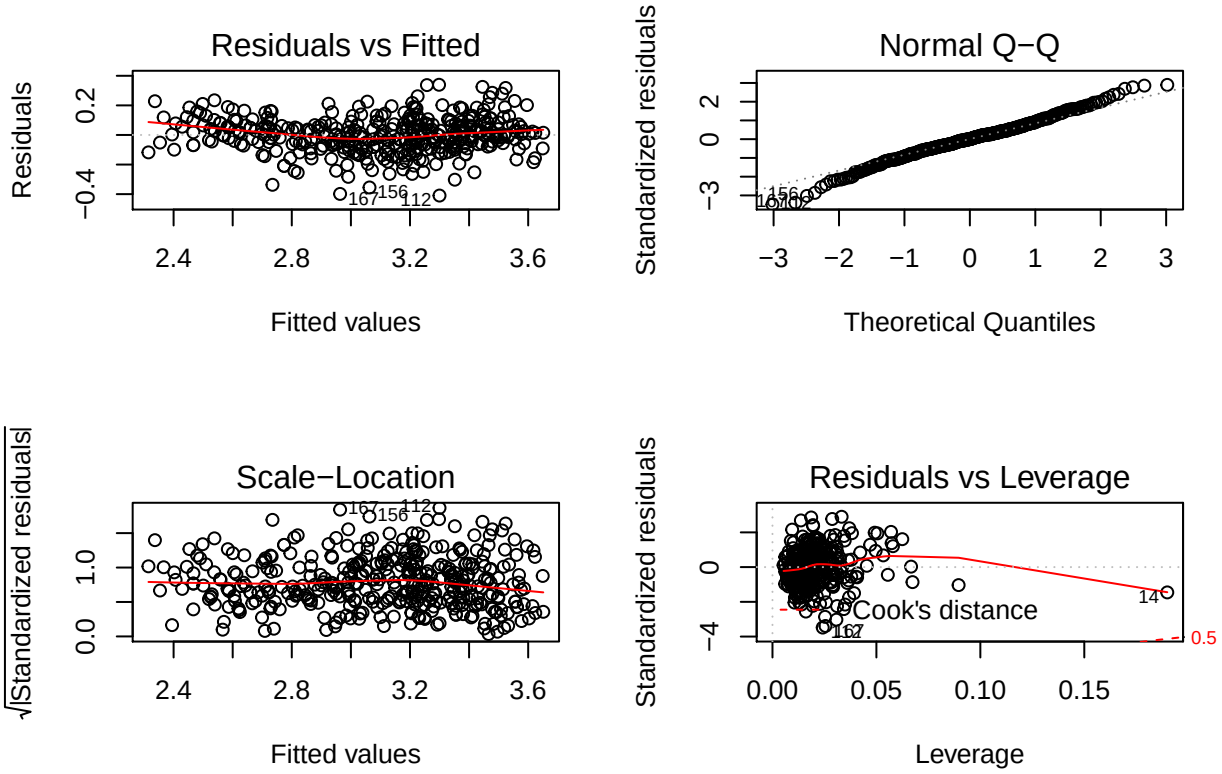
From the correlation matrix in 9a., displacement, horsepower and weight show a similar nonlinear pattern against our response mpg. This nonlinear pattern is very close to a log form.

```
lm.fit2<-lm(log(mpg)~cylinders+displacement+horsepower+weight+acceleration+year+origin,data=Auto)
summary(lm.fit2)
```

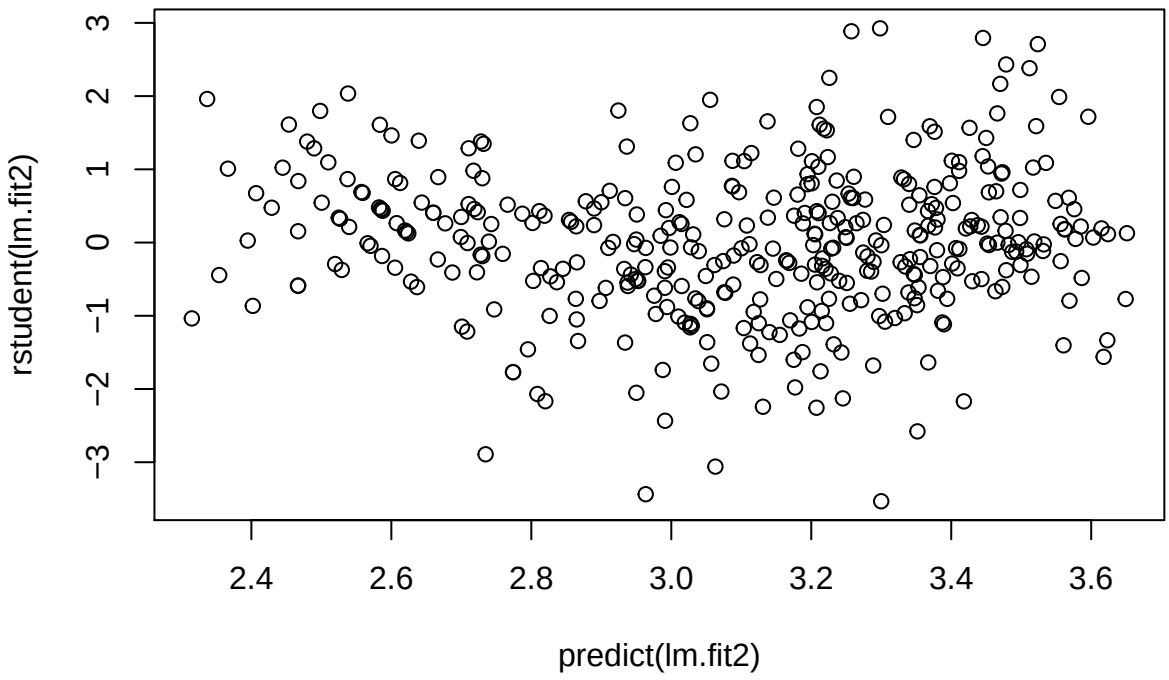
```
##
## Call:
## lm(formula = log(mpg) ~ cylinders + displacement + horsepower +
##     weight + acceleration + year + origin, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40955 -0.06533  0.00079  0.06785  0.33925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.751e+00  1.662e-01  10.533  < 2e-16 ***
## cylinders    -2.795e-02  1.157e-02  -2.415  0.01619 *
## displacement  6.362e-04  2.690e-04   2.365  0.01852 *
## horsepower   -1.475e-03  4.935e-04  -2.989  0.00298 **
## weight       -2.551e-04  2.334e-05 -10.931  < 2e-16 ***
## acceleration -1.348e-03  3.538e-03  -0.381  0.70339
```

```
## year          2.958e-02  1.824e-03  16.211  < 2e-16 ***
## origin        4.071e-02  9.955e-03   4.089 5.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1191 on 384 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8773
## F-statistic: 400.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm.fit2)
```




```
plot(predict(lm.fit2),rstudent(lm.fit2))
```



Chapter 3 10 (a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
summary(Carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75
##  Median :272.0   Median :117.0   Medium:219   Median :54.50
##  Mean   :264.8   Mean   :115.8                Mean   :53.32
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00
##  Max.   :509.0   Max.   :191.0                Max.   :80.00
##    Education    Urban       US
##  Min.   :10.0   No :118   No :142
##  1st Qu.:12.0   Yes:282   Yes:258
##  Median :14.0
##  Mean   :13.9
##  3rd Qu.:16.0
##  Max.   :18.0
```

```
attach(Carseats)
lm.fit = lm(Sales~Price+Urban+US)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## 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!

Price The linear regression suggests a relationship between price and sales given the low p-value of the t-statistic. The coefficient states a negative relationship between Price and Sales: as Price increases, Sales decreases.

UrbanYes The linear regression suggests that there isn't a relationship between the location of the store and the number of sales based on the high p-value of the t-statistic.

USYes The linear regression suggests there is a relationship between whether the store is in the US or not and the amount of sales. The coefficient states a positive relationship between USYes and Sales: if the store is in the US, the sales will increase by approximately 1201 units.

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

Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

(d) For which of the predictors can you reject the null hypothesi H0 : Bj = 0? Price and USYes, based on the p-values, F-statistic, and p-value of the F-statistic.

(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)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US)
##
## 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 the RSE and R^2 of the linear regressions, they both fit the data similarly, with linear regression from (e) fitting the data slightly better.

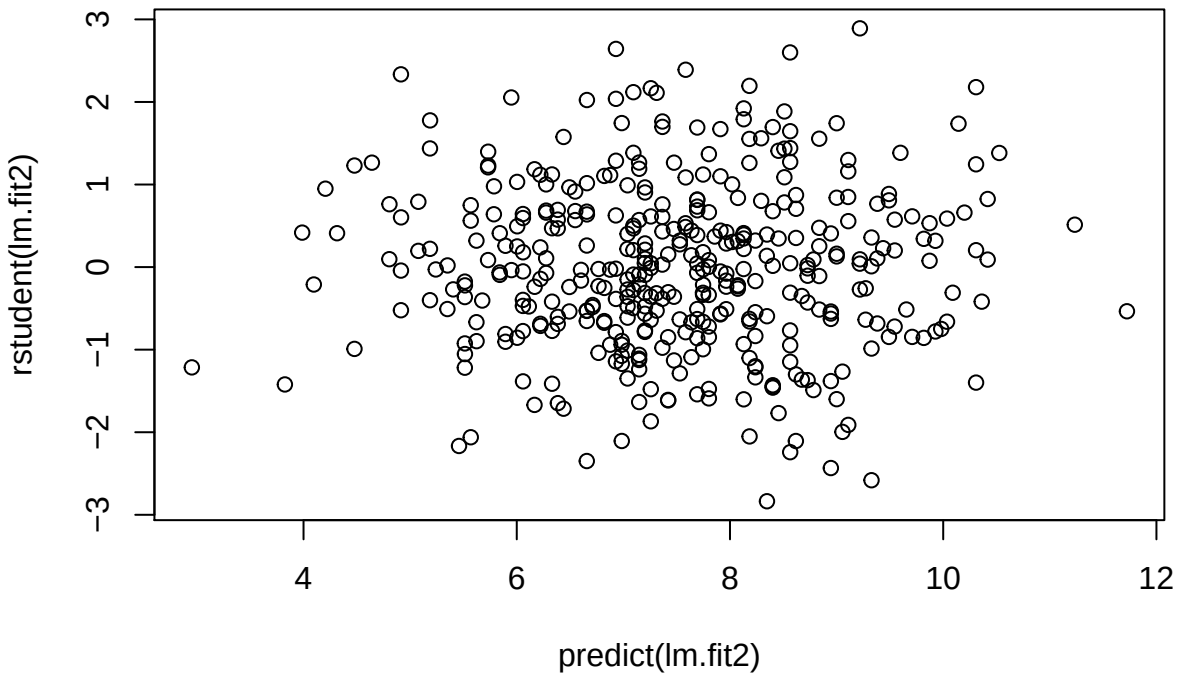
(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
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)?

```
plot(predict(lm.fit2), rstudent(lm.fit2))
```



Analysis of studentized residuals reveals no outliers. All the values are within range of -3 to 3.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.