HW2_STATS488

Melchor Ronquillo

2022-09-19

1) Chapter 3, Question 3

Suppose we have a data set with five predictors: X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, X5 = Interaction between GPA and Level.

The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get * B0 = 50,B 1 = 20, B2 = 0.07, B3 = 35, B4 = 0.01, B5 = 10

```
a) which answer is correct, and why?

    Y = 50 + 20X1(GPA) + 0.07X2(IQ) + 35X3(Level) + 0.01X4(GPA:IQ) - 10X5(GPA:Level)

    For level = college(1)
    Y = 50 + 20X1(GPA) + 0.07X2(IQ) + 35(1) + 0.01X4(GPA*IQ) - 10X5(GPA(1))
    Y = 85 + 20X1(GPA) + 0.07X2(IQ) + 0.01X4(GPA*IQ) - 10(GPA)
    Y = 85 + 10X1(GPA) + 0.07X2(IQ) + 0.01X4(GPA*IQ)

    For level = high school(0):
    Y = 50 + 20X1(GPA) + 0.07X2(IQ) + 35(0) + 0.01X4(GPA*IQ) - 10X5(GPA(0))
    Y = 50 + 20X1(GPA) + 0.07X2(IQ) + 0.01X4(GPA*IQ)

    iii. For a fixed value of IQ and GPA, high school graduates earn more,
    on average, than college graduates provided that the GPA is high enough
    with IQ and GPA being the same fixed value, each equation differs as so:

    college = Y = 85 + 10X1(GPA)
    high school = Y = 50 + 20X1(GPA)

    answers i and ii cannot be correct because it there IS a possibility that one
    group could earn more than the other IF a specific condition is met, which in
    this case is GPA. It may appear that college graduates earn more than high school
    graduates based on the equation with their B0 as 85 vs 50 but if the GPA is at
    least 3.5 or higher then high schoolers will on average earn more than college graduates.

b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

    Y = 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(440) - 1(4)
```

```r
Y = 50 + 80.0 + 77 + 35 + 44 - 4
Y
```

```
## [1] 282
```

$282,000

```
c) True or false: Since the coefficient for the GPA/IQ interaction term
is very small,there is very little evidence of an interaction effect.
Justify your answer.
    Cannot conclude the significance of interaction between terms based
    on it's coefficients.The significance of the interaction between two
    variables can be determined by the P-Value and base it off from the
    significance level.
```

2) Chapter 3, Problem 9

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

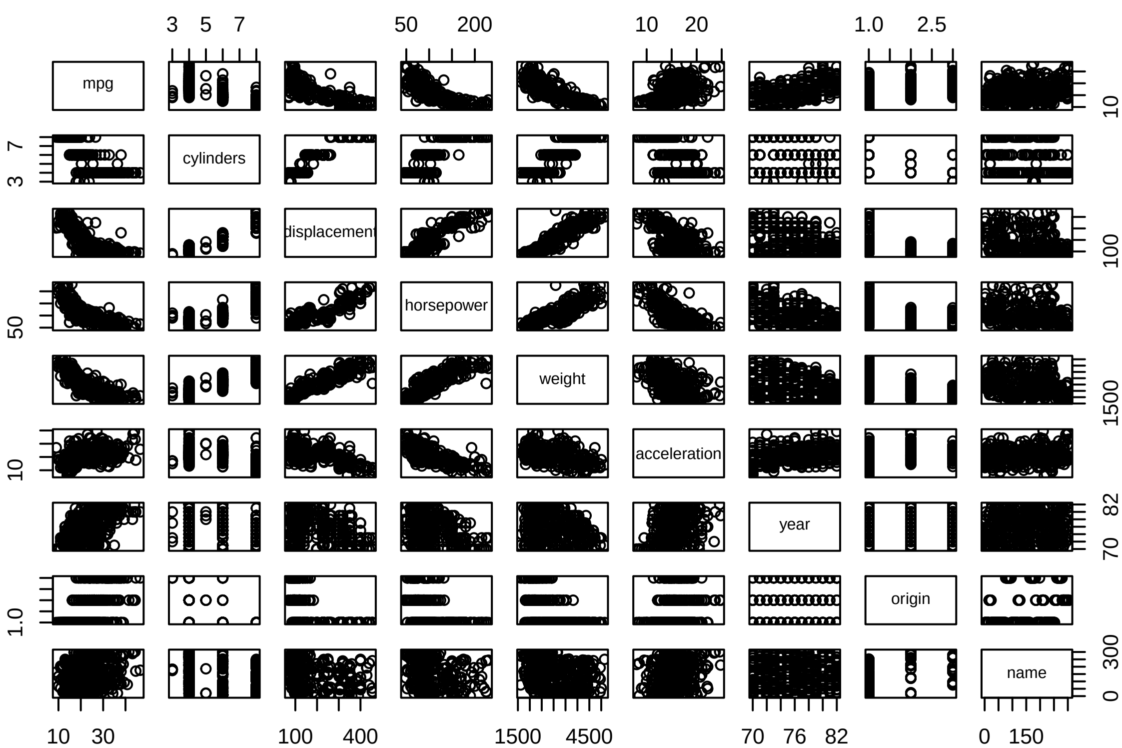
```r
install.packages("ISLR",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/b4/vbzhgztj3tj299xgxnklp28c0000gn/T//RtmpIuknOt/downloaded_packages
```

```r
library('ISLR')
```

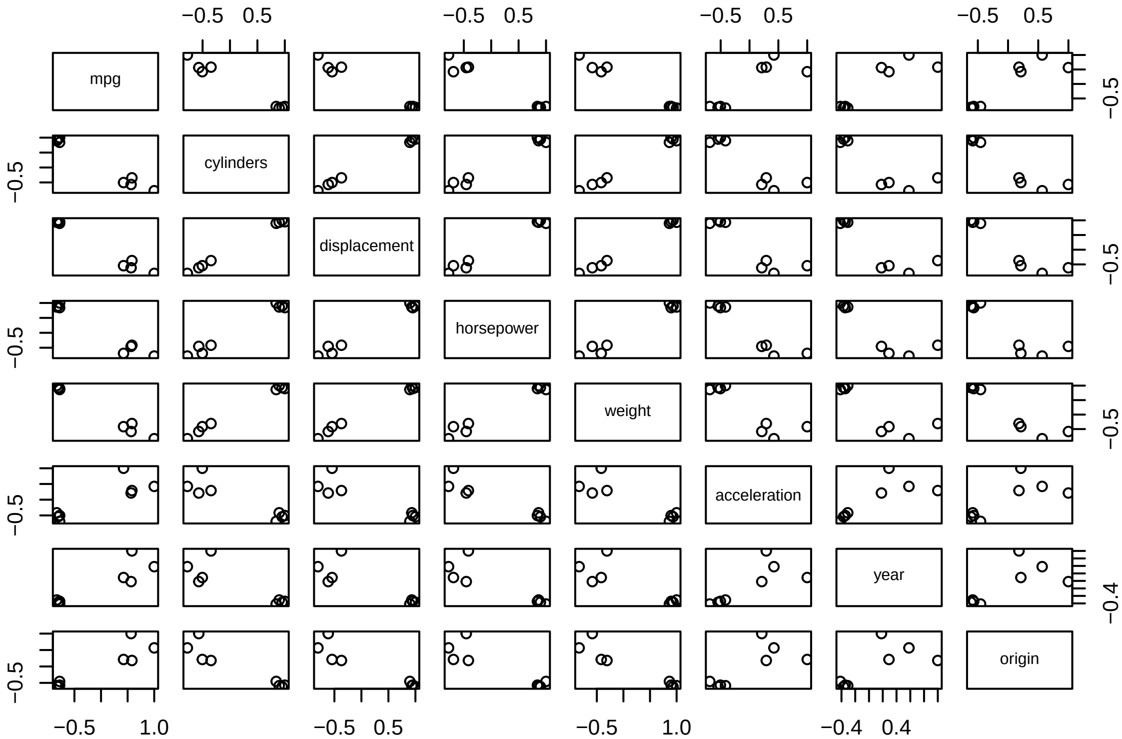
```
a) Produce a scatterplot matrix which includes all of the variables in
the data set.
```

```r
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.
```

```r
Auto_Noname = Auto[, c("mpg", "cylinders",  "displacement", "horsepower",
                       "weight", "acceleration", "year", "origin")]
Auto_cor = cor(Auto_Noname)
pairs(Auto_cor)
```



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.

```
mpg.lm = lm(mpg~., data = Auto_Noname)
summary(mpg.lm)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = Auto_Noname)
##
## 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
```

Comment on the output. For instance:

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

Relationship with the predictors and response can be measured by the significance of each p - value. some predictiors appear to have a significant relationship with the response while others do not.

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

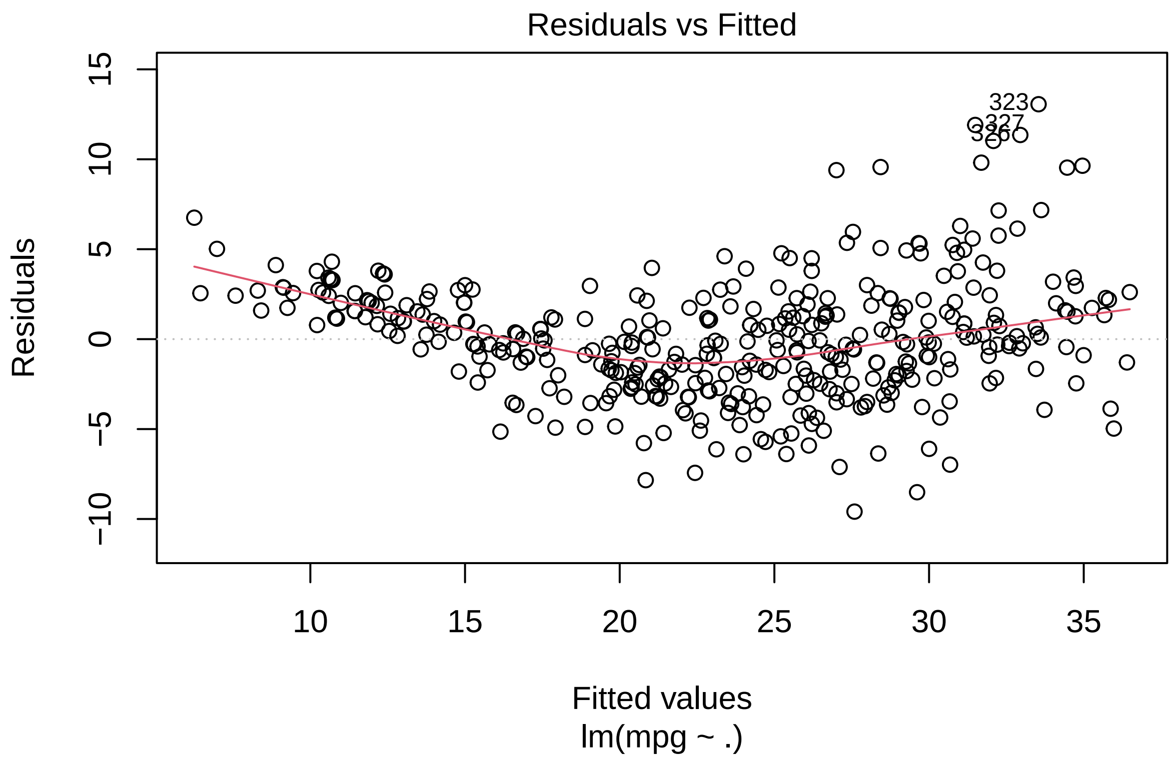
We can determine whether a predictor appears to a have statistically significant relationship to the response based on the significance code given in the summary. The number of *s next to a specific variable presents how significant a variable is to the response based on it's p-value. The hypothesis is H0: B = 0, and if a P-Value is less than the significance value (usually .05 for 95% confidence interval), then the null hypothesis is rejected and it shows that there is a non zero correlation between the predictor and response.In this regression with mpg as the response, the predictors weight, year, and origin have ***, meaning that it is statistically significant at a 100% confidence interval. Displacement is also statistically significant with ** meaning it is significant at a 99.9% confidence interval.

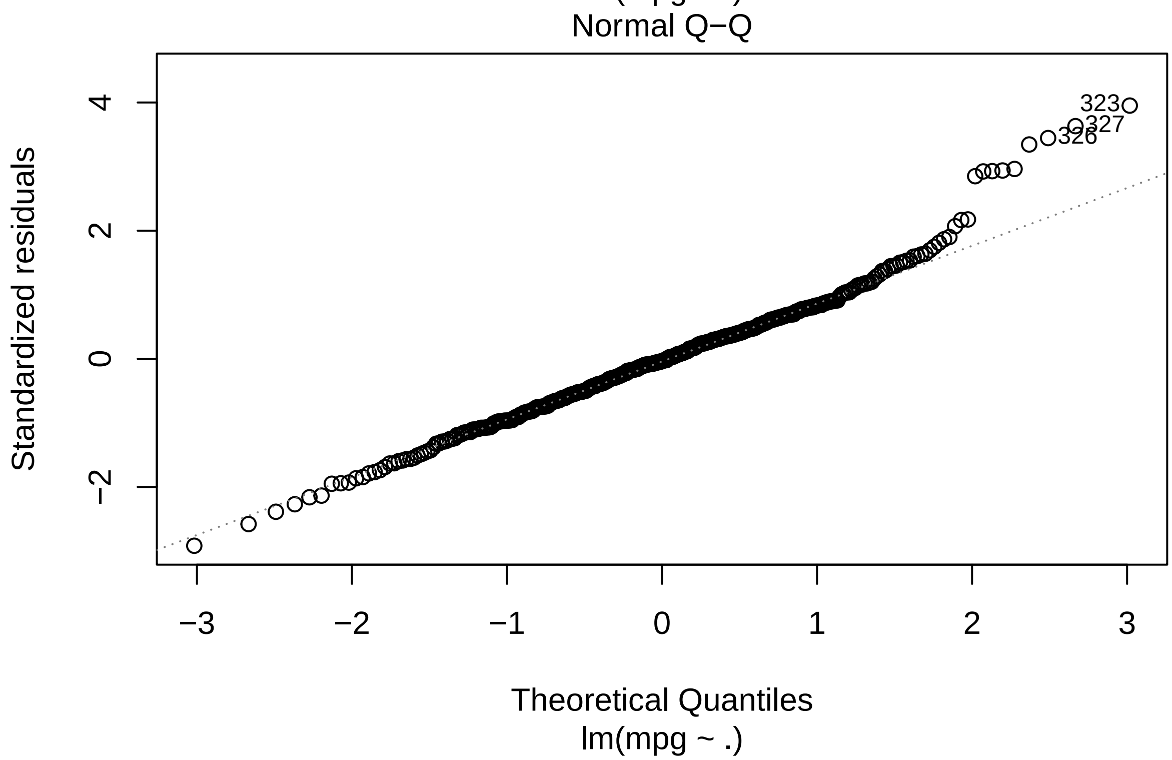
iii. What does the coefficient for the year variable suggest?

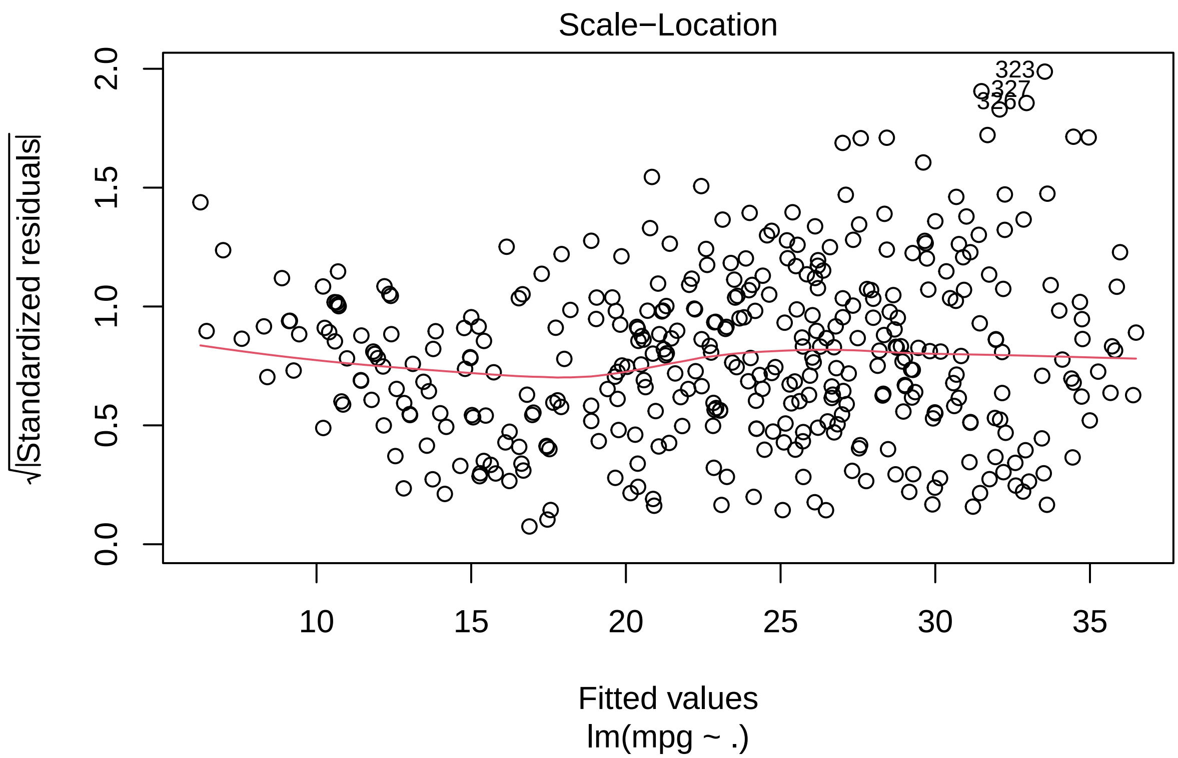
The predictor year has a coefficient of 0.750773. For every increment of 1 that year increases, the mpg will go up by 0.750773

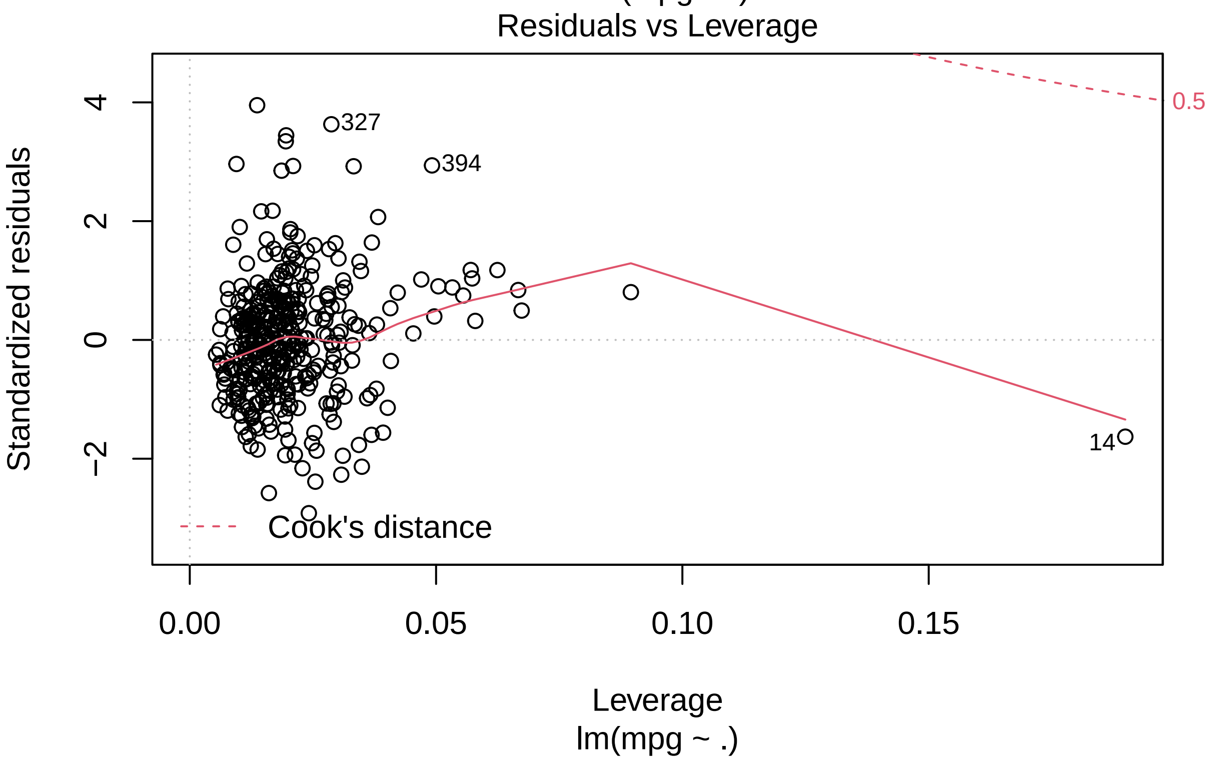
d) Use the plot() function to produce diagnostic plots of the linear regression fit.

```
plot(mpg.lm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ .)
323
327
326


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ .)
323
327
326






```
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?
```

Based on the plots, there do not appear to be any unusually large outliers.

```
There are some points that stray away from the line in teh Normal QQ
plot, but since all plots seem to remain within -2 to 2 in the
scale location plot and there are no points that exceed the dotted red line
in the cooks distance plot, the fit does not seem to have any observations with
unusually high outliers and leverage.
```

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

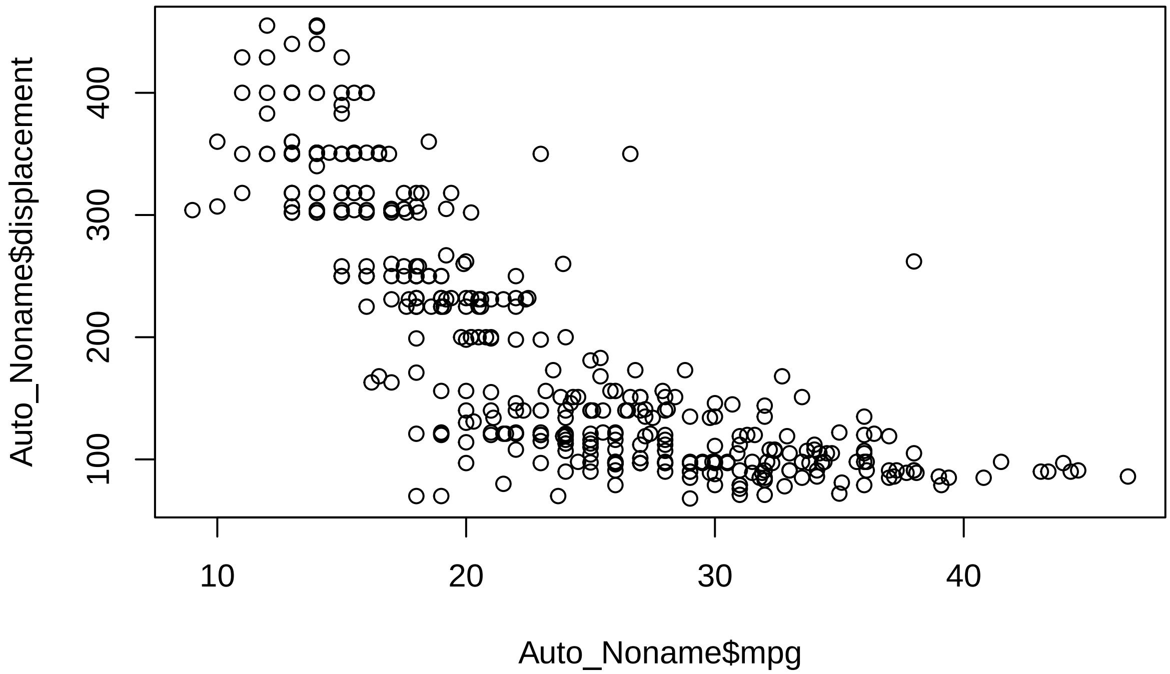
```r
mpg.lm2 = lm(mpg~.+ displacement*horsepower + weight*horsepower +
               displacement*cylinders, data = Auto_Noname)
summary(mpg.lm2)
```

```
##
## Call:
## lm(formula = mpg ~ . + displacement * horsepower + weight * horsepower +
##     displacement * cylinders, data = Auto_Noname)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8472 -1.5513 -0.0656  1.3490 12.0143
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.894e+00  4.566e+00   0.415  0.67858
## cylinders               1.321e-01  5.683e-01   0.232  0.81633
## displacement           -4.975e-02  1.886e-02  -2.638  0.00868 **
## horsepower             -2.178e-01  2.680e-02  -8.125 6.33e-15 ***
## weight                 -6.590e-03  1.559e-03  -4.228 2.95e-05 ***
## acceleration           -1.636e-01  9.404e-02  -1.739  0.08277 .
## year                    7.527e-01  4.477e-02  16.813  < 2e-16 ***
## origin                  6.725e-01  2.570e-01   2.616  0.00924 **
## displacement:horsepower  2.954e-04  1.041e-04   2.837  0.00480 **
## horsepower:weight       2.472e-05  1.043e-05   2.370  0.01829 *
## cylinders:displacement   1.390e-03  2.325e-03   0.598  0.55023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.896 on 381 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8623
## F-statistic: 245.9 on 10 and 381 DF,  p-value: < 2.2e-16
```

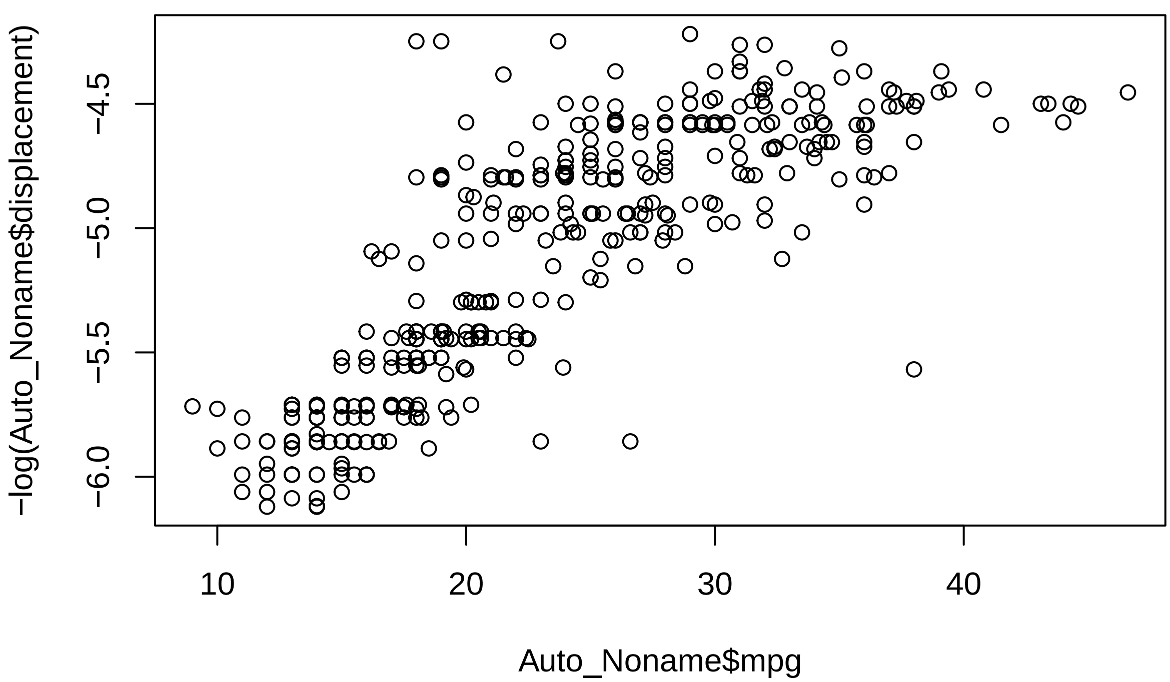
```
Based on the interactions, displacement:power and horsepower:weight
both appear to be statistically significant. This makes sense because
displacement and horsepower relates in the sense that a vehicle with
a hoigh displacement has a bigger engine ann combined with high horsepower
means that it is either a fast performance car or a big truck, both vehicles
that are known for a lower mpg. Horsepower and weight also interact in
a sence that a lighter car with high horsepower will still have
better mpg than a heavy car with the same amount of horsepower because
it will take more gas to get the heavy car going.
```

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

```
plot(Auto_Noname$mpg, Auto_Noname$displacement)
```




```
plot(Auto_Noname$mpg, -log(Auto_Noname$displacement))
```



By taking the -log of displacement, I was able to transform the plot of mpg and displacement to a more linear relationship. This will also work

```
with horsepower and weight.
```

3) Chapter 3, Problem 10

This question should be answered using the Carseats data set.

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

```
sales.mlr = lm(Sales~Price+Urban+US, data = Carseats)
summary(sales.mlr)
```

```
##
## 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!

    For every dollar increase in price, sales will decrease by 0.054459
    thousand. If the carseat is in an urban area, sales will decrease by
    0.021916 thousand. If the carseat is in the US, sales will go up by
    1.200573 thousand


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

    Sales = 13.043469  - 0.054459X1(Price) - 0.021916(Urban) + 1.200573(US)

d) For which of the predictors can you reject the null hypothesis
H0:B=0?

  We can reject null hypothesis for Price and US, the p-value of both
  variables are less than 0.05. 95% confident that those predictors
  each have a non zero correlation with Sales.

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.

```r
sales.mlr2 = lm(Sales~Price+US, data = Carseats)
summary(sales.mlr2)
```

```
##
## 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?

Both models fit the model similarly well, with no major difference between the R-squared values. This means that Urban had no real impact in prediciting the sales of carseats.

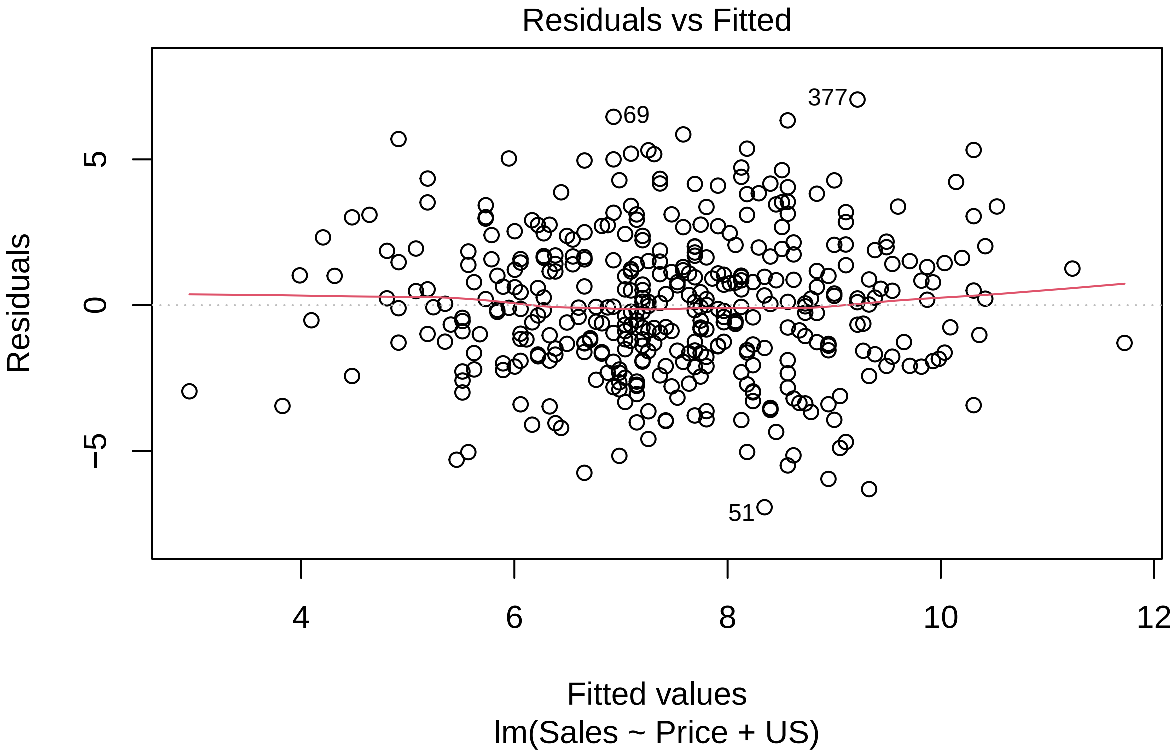
g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

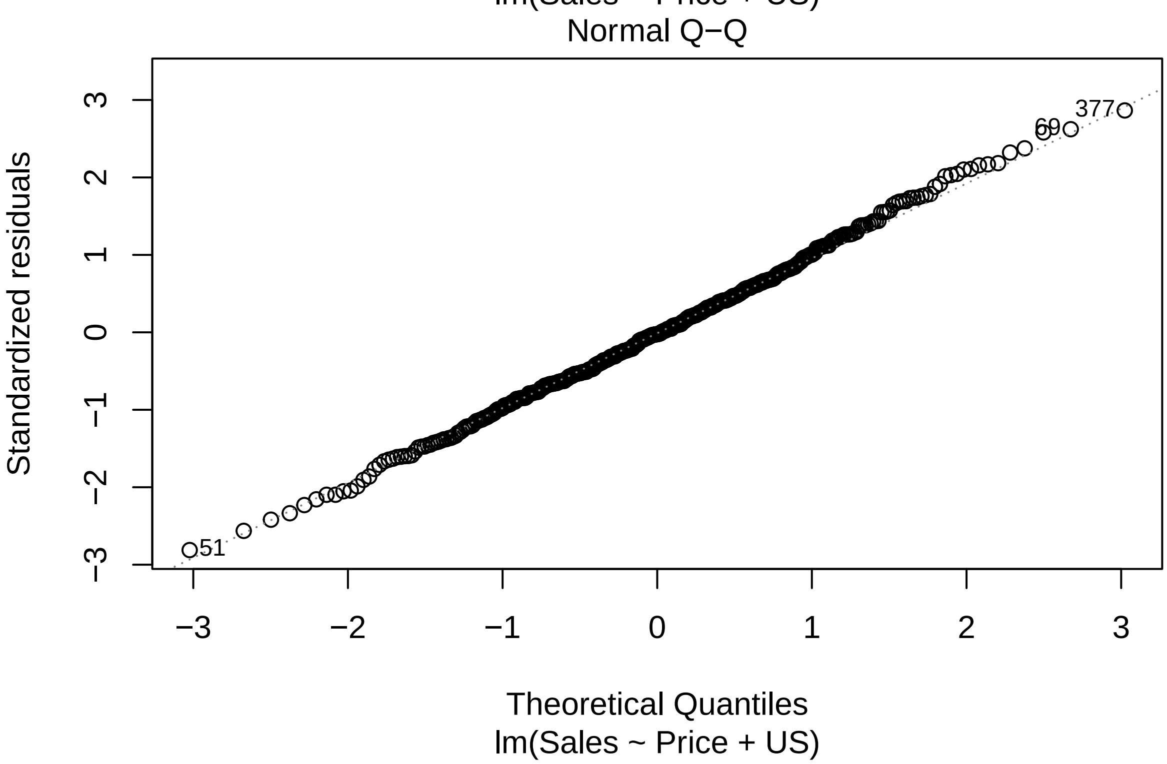
```r
confint(sales.mlr2)
```

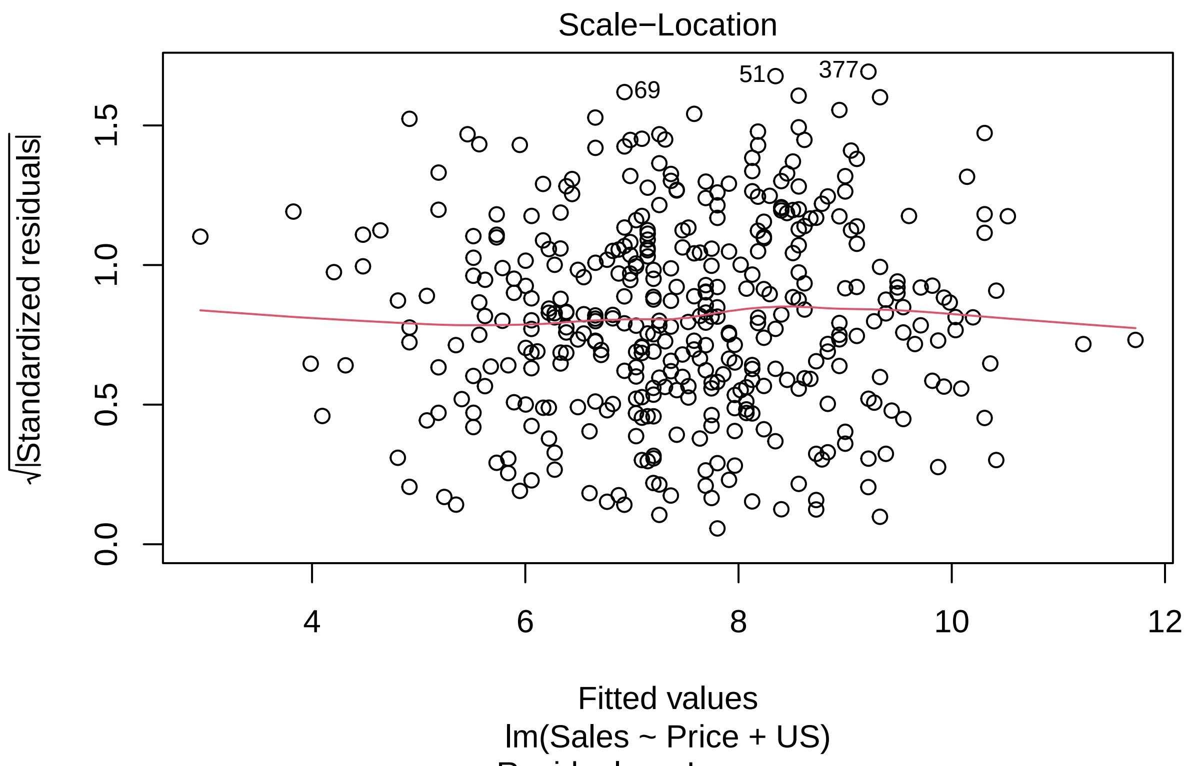
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

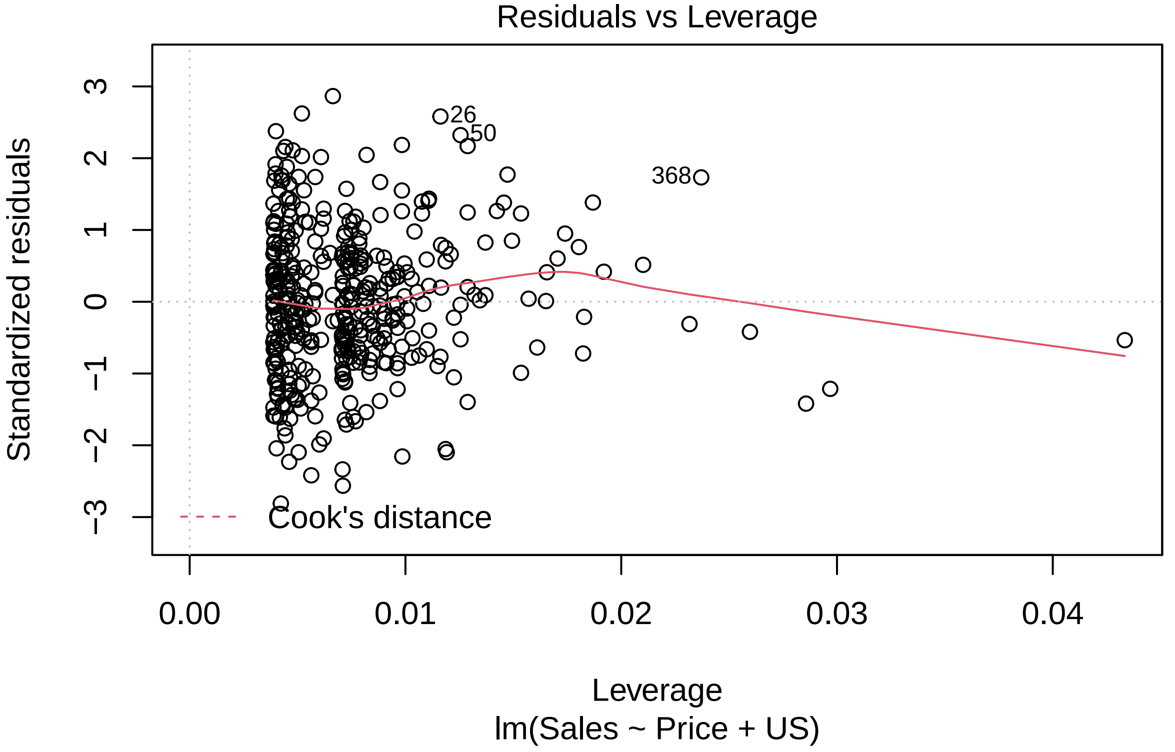
h) Is there evidence of outliers or high leverage observations in the model from (e)?

```r
plot(sales.mlr2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Sales ~ Price + US)
69
377
51


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ Price + US)
51
69
377






```
No evidence of outliers or high leverage obersavtions in model
```

4) Chapter 4, question 6

Suppose we collect data for a group of students in a statistics class with variables: X1 = hours studied, X2 = undergrad GPA, Y = receive an A. We fit a logistic regression and produce estimated coefficient: B0 = -6, B1 = 0.05, B2 = 1.

Y = -6 + 0.05X1(Hours) + 1X2(GPA)

```
a) Estimate the probability that a student who studies for 40 h and has an
undergrad GPA of 3.5 gets an A in the class.

    p(x) = (e^(B0 + B1X1 + B2X2)) / (1 + e^(B0 + B1X1 + B2X2))
```

```r
Py <- (exp(-6 + (0.05*40) + 3.5) / (1 + exp(-6 + (0.05*40) + 3.5)))
Py
```

```
## [1] 0.3775407
```

```
    The probability that a student who studies for 40 hours and has a
    GPA of 3.5 gets an A is 37.75%


b) How many hours would the student in part (a) need to study to have a
50 % chance of getting an A in the class?

    0.50 = (e^(-6 + 0.05X1 + 3.5)) / (1 + e^(-6 + 0.05X1 + 3.5)),
    solve for X1

    log(p(x) / 1 - p(x)) = B0 + B1X1 + B2X2
    log(0.50 / 1 - 0.50) = -6 + 0.05X1 + 3.5
    log(0.50 / 1 - 0.50) + 6 - 3.5 = 0.05X1
    (log(0.50 / 1 - 0.50) + 6 - 3.5 ) / 0.05 = X1
```

```r
x <- (0.50 / (1 - 0.50))
x
```

```
## [1] 1
```

```r
logg <- log(x)
logg
```

```
## [1] 0
```

```r
X1 = (logg + 6 - 3.5 ) / 0.05
X1
```

```
## [1] 50
```

```
    A student with a GPA of 3.5 would have to study for 50 hourse to have
    a 50% chance of getting an Ain the class
```

5) Chapter 4, question 16

Using the Boston data set, fit classification models in order to predict whether a given census tract has a crime rate above or below the median. Explore logistic regression, LDA, naive Bayes, and KNN models using various subsets of the predictors. Describe your findings.

```r
library(MASS)
library(class)
#Boston
```

```
for every row with crim > 2.5, classify as 1, else 0
```

```r
Boston$I_crime <- (Boston$crim > median(Boston$crim)) + 0
#Boston
```

Remove original crim column

```r
Boston = subset(Boston, select = -c(crim))
#Boston
```

Split data into training (70%) and testing (30%)

```r
set.seed(338)
u <- runif(nrow(Boston))
#u
train <- Boston[u <= 0.7,]
test <- Boston[u >  0.7,]
#train
```

Create logistic regression

```r
crimlog <- glm(I_crime~ ., data = train, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Predict probabilities

```r
crimlog.probs <- predict(crimlog, test, type = "response")
crimlog.probs
```

```
##            3            5            9           17           20           23
## 5.597437e-02 2.277103e-01 2.717790e-01 4.233167e-01 3.975784e-01 5.372176e-01
##           24           27           31           33           35           40
## 5.914278e-01 7.955392e-01 8.481375e-01 9.994343e-01 9.988630e-01 3.058484e-06
##           43           44           47           48           54           55
## 6.098383e-03 3.717289e-03 2.715824e-03 1.695248e-02 8.550819e-04 1.598654e-06
##           58           61           65           71           77           80
## 2.025186e-06 1.133222e-01 6.196133e-02 1.557259e-03 1.015839e-02 1.379003e-03
##           82           85           93          100          101          102
## 1.344521e-03 8.204556e-03 3.995880e-04 7.238342e-03 4.677627e-01 3.711155e-01
##          103          109          111          112          116          118
## 9.999987e-01 2.866403e-01 1.701267e-01 3.561241e-01 7.449963e-01 2.238657e-01
##          122          126          128          133          137          147
## 3.843937e-01 3.385317e-01 7.580644e-01 9.405905e-01 8.893383e-01 1.000000e+00
##          151          153          154          155          158          159
## 9.999949e-01 9.999978e-01 1.000000e+00 9.999997e-01 9.629523e-01 7.611854e-01
##          160          161          162          163          166          172
## 9.999974e-01 9.605555e-01 9.799838e-01 9.911013e-01 9.988742e-01 7.789021e-01
##          176          179          181          182          185          193
## 8.401879e-02 2.366772e-01 3.558871e-01 8.142617e-02 7.760377e-02 2.575910e-04
##          194          200          201          203          210          213
## 4.818796e-07 2.246499e-07 3.175719e-07 4.973553e-07 2.138644e-01 9.967422e-02
##          219          220          223          226          228          230
## 6.069900e-01 7.134192e-01 9.010221e-01 9.913693e-01 9.096984e-01 5.192880e-01
##          244          245          248          251          252          257
## 1.314168e-03 7.606999e-02 1.225598e-01 9.415136e-03 1.874436e-02 1.644594e-06
##          263          264          266          269          271          275
## 9.938581e-01 8.971483e-01 3.570157e-01 4.475512e-01 1.473891e-03 5.719079e-04
##          279          280          281          285          292          293
## 1.282715e-04 1.710294e-03 5.067241e-02 6.901154e-08 6.133620e-06 1.341427e-06
```

```
##          302          306          307          317          318          321
## 4.711690e-04 2.058555e-02 5.045615e-02 3.818340e-01 2.778942e-01 1.743585e-01
##          323          324          325          328          332          339
## 1.098875e-01 1.986207e-01 1.602345e-01 2.115473e-01 9.794205e-06 3.894397e-01
##          340          341          346          350          351          353
## 3.880305e-01 4.567623e-01 1.857724e-02 3.813129e-04 1.822847e-04 2.920412e-05
##          355          356          357          365          375          376
## 4.737576e-05 8.379896e-05 1.000000e+00 1.000000e+00 9.999995e-01 9.999998e-01
##          377          380          385          389          392          393
## 9.999999e-01 9.999996e-01 1.000000e+00 1.000000e+00 1.000000e+00 9.999998e-01
##          397          398          399          401          404          407
## 9.999999e-01 9.999998e-01 9.999997e-01 9.999998e-01 9.999997e-01 9.999995e-01
##          410          411          419          423          434          436
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          439          440          441          447          449          450
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          455          458          460          461          462          463
## 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##          465          467          468          474          478          483
## 9.999997e-01 1.000000e+00 9.999998e-01 9.999998e-01 9.999995e-01 9.999781e-01
##          487          489          490          498          503
## 9.999962e-01 6.816888e-02 2.698968e-01 7.285586e-01 1.691067e-01
```

Where probablity is atleast .50, classify as 1, else 0

```
crimlog.pred <- (crimlog.probs >= .5) + 0
crimlog.pred
```

```
##   3   5   9  17  20  23  24  27  31  33  35  40  43  44  47  48  54  55  58  61
##   0   0   0   0   0   1   1   1   1   1   1   0   0   0   0   0   0   0   0   0
##  65  71  77  80  82  85  93 100 101 102 103 109 111 112 116 118 122 126 128 133
##   0   0   0   0   0   0   0   0   0   0   1   0   0   0   1   0   0   0   1   1
## 137 147 151 153 154 155 158 159 160 161 162 163 166 172 176 179 181 182 185 193
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   0   0   0   0   0   0
## 194 200 201 203 210 213 219 220 223 226 228 230 244 245 248 251 252 257 263 264
##   0   0   0   0   0   0   1   1   1   1   1   1   0   0   0   0   0   0   1   1
## 266 269 271 275 279 280 281 285 292 293 302 306 307 317 318 321 323 324 325 328
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## 332 339 340 341 346 350 351 353 355 356 357 365 375 376 377 380 385 389 392 393
##   0   0   0   0   0   0   0   0   0   0   1   1   1   1   1   1   1   1   1   1
## 397 398 399 401 404 407 410 411 419 423 434 436 439 440 441 447 449 450 455 458
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 460 461 462 463 465 467 468 474 478 483 487 489 490 498 503
##   1   1   1   1   1   1   1   1   1   1   1   0   0   1   0
```

```
#length(crimlog.pred)
```

create confusion matrix and show error rate

```
table(crimlog.pred, test$I_crime) #Confusion Matrix
```

```
##
## crimlog.pred  0  1
##            0 71 10
##            1  4 70
```

```r
(10+4) / length(test$I_crime) #Error rate
```

```
## [1] 0.09032258
```

LDA

```r
crimlda <- lda(I_crime~., data = train, family = "binomial")
crimlda.pred = predict(crimlda, test)
table(crimlda.pred$class, test$I_crime) #Confusion Matrix
```

```
##
##      0  1
##   0 72 19
##   1  3 61
```

```r
(3+19) / length(test$I_crime) #Error rate
```

```
## [1] 0.1419355
```

KNN

```r
library(carData)
library(class)
crimknn <- knn(train = train[,1:13], test = test[,1:13], cl = train[,14], k = 1)
table(test$I_crime, crimknn) #Confusion Matrix
```

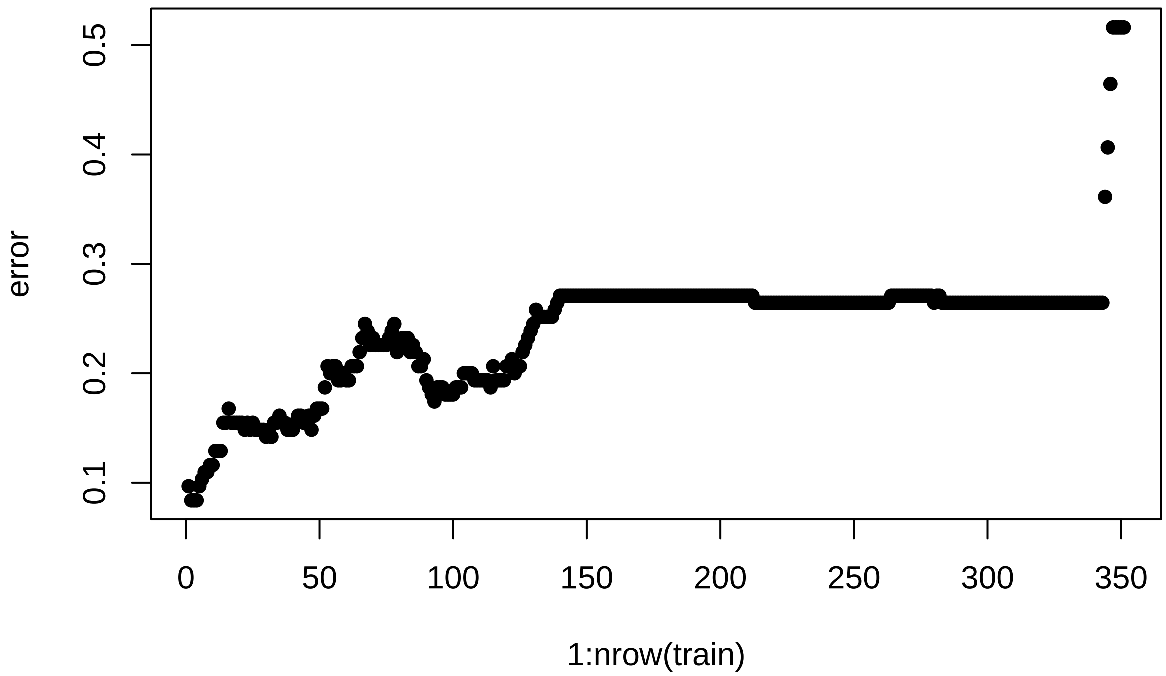
```
##    crimknn
##      0  1
##   0 69  6
##   1  9 71
```

```r
(6+9) / length(test$I_crime) #Error rate
```

```
## [1] 0.09677419
```

```
I want to try and find a better K by plotting the errors and finding the
lowesr point as my K
```

```r
k = 1
error <- c()
for (k in 1:nrow(train)){
crimknn2 <- knn(train = train[,1:13], test = test[,1:13], cl = train[,14], k = k)
error[k] <- mean(test$I_crime != crimknn2)
}
#View(error)
plot(1:nrow(train), error, pch = 16) #Lowest points are closest to x = 0, try k = 5
```




```r
crimknn <- knn(train = train[,1:13], test = test[,1:13], cl = train[,14], k = 10)
table(test$I_crime, crimknn) #Confusion Matrix
```

```
##    crimknn
##      0  1
##   0 67  8
##   1 10 70
```

```r
(8+10) / length(test$I_crime) #Error rate
```

```
## [1] 0.116129
```

Turns out, K = 1 yielded beter results

Naive Bayes

```r
install.packages("e1071", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/b4/vbzhgztj3tj299xgxnklp28c0000gn/T//RtmpIukn0t/downloaded_packages
```

```r
library(e1071)
crimnb <- naiveBayes(I_crime~., data = train)
crimnb.pred <- predict(crimnb, test)
table(crimnb.pred, test$I_crime) #Confusion Matrix
```

```
##
## crimnb.pred  0  1
##           0 67 18
##           1  8 62
```

```r
(18+8) / length(test$I_crime) #Error rate
```

```
## [1] 0.1677419
```

Out of all of the different models for this dataset, my logistic regression

predicted whether a given census tract has a crime rate above or below the median the best, as it had the lowest error/ missclassification rate of 9.03%.
KNN came in 2nd with 9.67%, LDA with 14.19%, and Naive Bayes with 16.77%