Assignment #2

Tina Roha

2/26/2021

Linear Regression

Chapter 3 (page 120): Questions 2, 9, 10, and 12

Question-2

Carefully explain the differences between the KNN classifier and KNN regression methods.

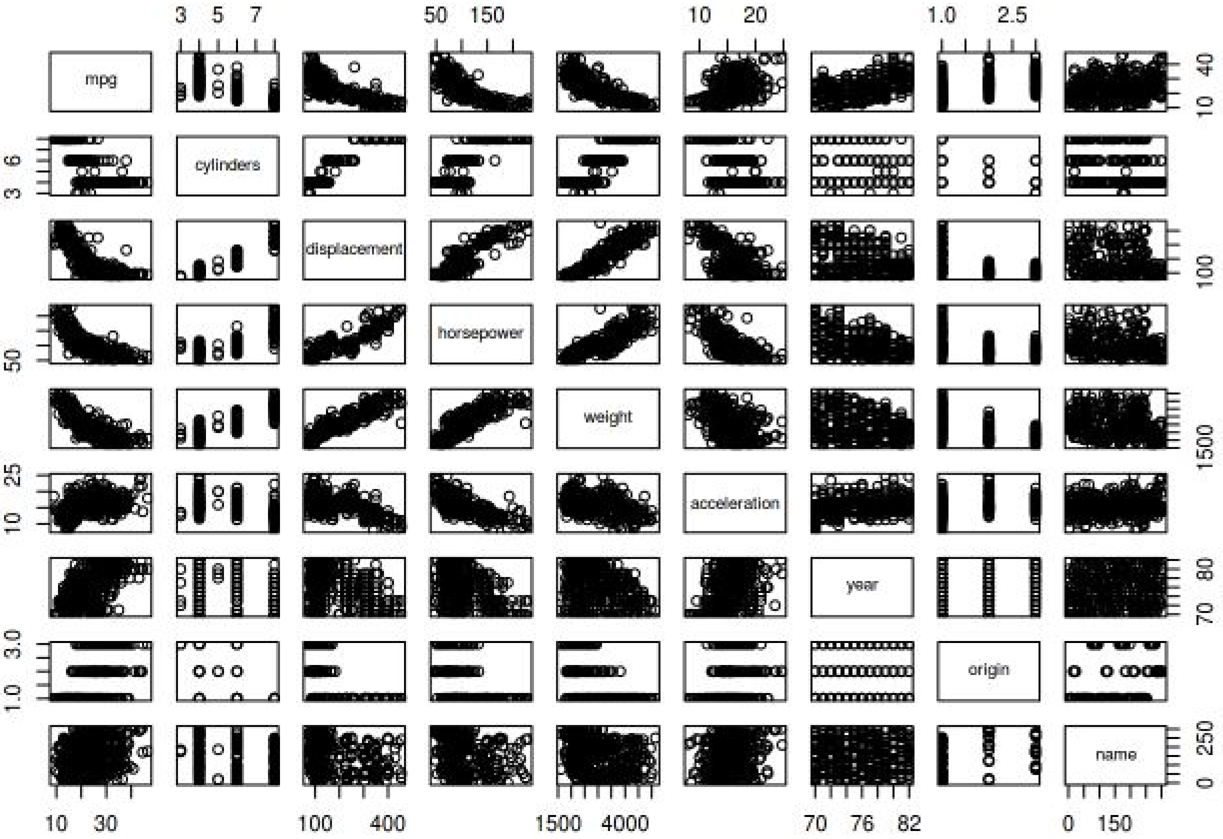
The KNN classifier is utilized to find solutions for problems with a qualitative response by establishing what the neighborhood of x_0 identifies as and then estimates the conditional probability $P(Y = j \mid X = x_0)$ for class j as the fraction of points in the neighborhood whose response values equal j. On the other hand, KNN regression methods is utilized to find solutions for problems with a quantitative response by establishing what the neighborhood of x_0 identifies as and then estimates the $f(x_0)$ as the mean of all the training responses in the neighborhood.

Question-9

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.

```
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.

```
names(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"  
## [5] "weight"       "acceleration" "year"         "origin"      
## [9] "name"
```

```
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. 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:

```r
fit2 <- lm(mpg ~ . - name, data = Auto)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

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

In this case, a relationship between the predictors and the response does exist. Furthermore, the output shows the p-value for the F-statistic is 2.2e-16, which is less than 0.05, indicating significance as well as a relationship between mpg and the rest of the variables.

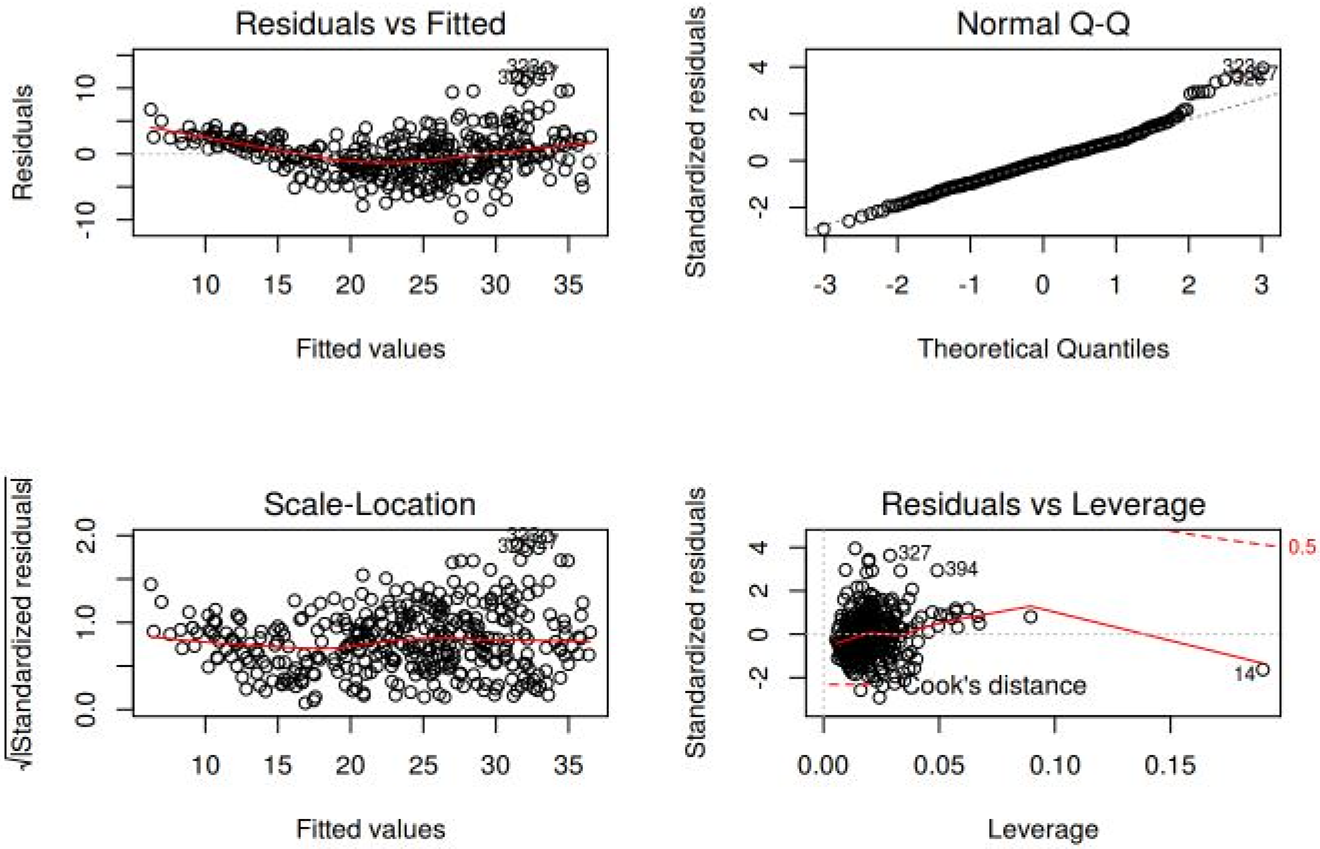
ii. Which predictors appear to have a statistically significant relationship to the response?

The predictors that appear to have a statistically significant relationship to the response includes displacement, weight, year and origin.

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

The coefficient for the year variable suggests, with all other predictors staying constant, in approximately one year mpg will have an increase of 0.7507727. Moreover, this concludes that vehicles are more fuel efficient by about one mpg each year.

d. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?



In this case, there seems to be some non linearity in the data due to the fact that some residuals, in the Residuals vs Fitted plot, stray far from the line. The Residual vs Leverage plot shows a high leverage point at point 14.

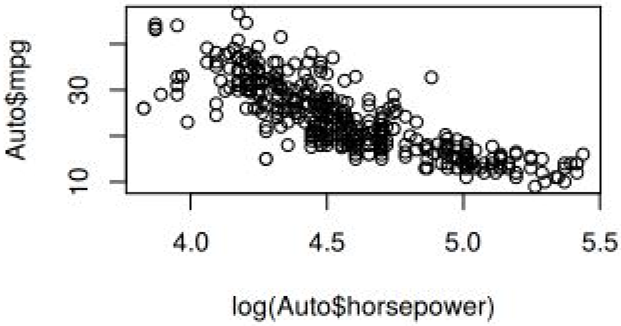
e. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

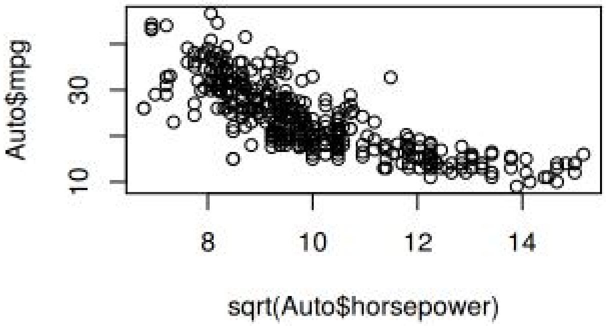
```
fit3 <- lm(mpg ~ cylinders * displacement+displacement * weight, data = Auto[, 1:8])
summary(fit3)
```

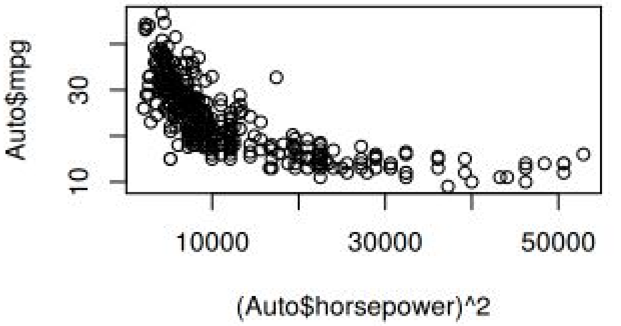
```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto[, 1:8])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                          Estimate Std. Error t value 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
```

The results conclude that the p-value for displacement:weight is significant with a value of 2.64e-05. Additionally, it can be concluded that the p-value for cylinders:displacement is not significant with a value of 0.384 due to the fact it is greater than the desired 0.05 or below requirement.

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





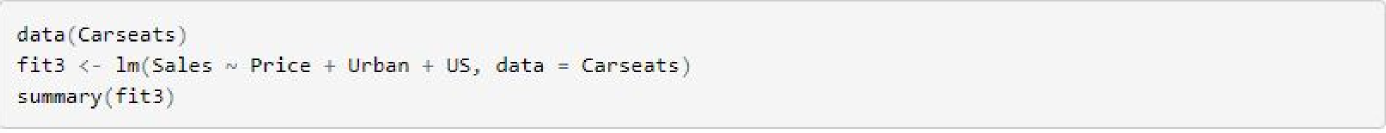


After applying logX, sqrtX, and X^2 to the horsepower predictor, the plot showing the best linearity is the logX transformation.

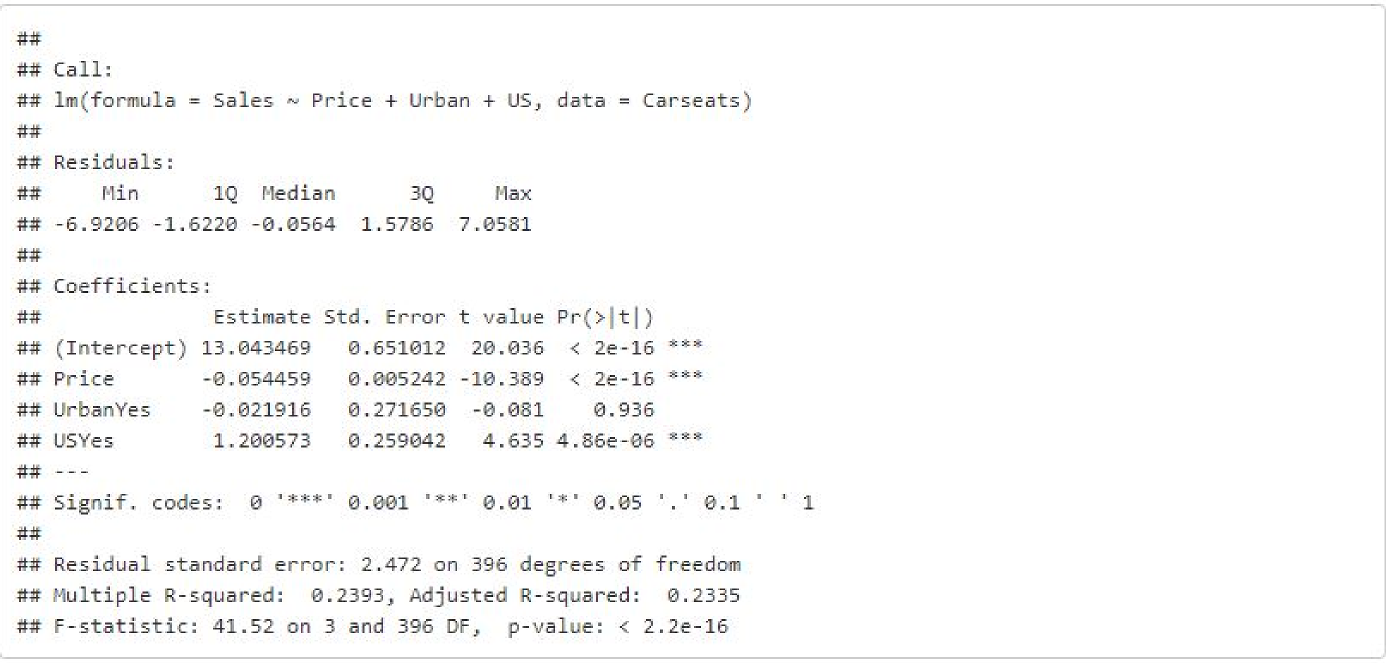
Question-10

This question should be answered using the Carseats data set.

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



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



```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b. Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

In the model, the interpretation of the Price variable coefficient is that by raising prices by one dollar it will cause a decrease in sales by approximately 54.4459 units. In addition, the interpretation of the Urban variable coefficient is that the sales of units in urban areas are approximately 21.916 units less when compared to rural areas. Additionally, the interpretation of the US variable coefficient is that the unit sales in a US store are approximately 1200.573 units greater when compared to other stores located outside of the US. In all three interpretations all other predictors remain fixed.

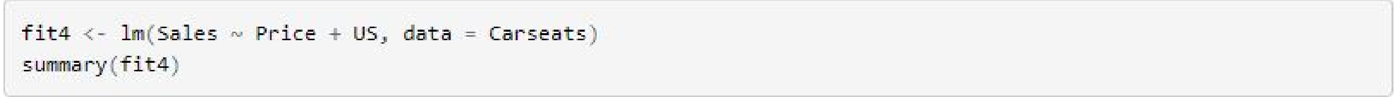
c. Write out the model in equation form, being careful to handle the qualitative variables properly.

Sales=13.0434689+(−0.0544588)×Price+(−0.0219162)×Urban+(1.2005727)×US+ε

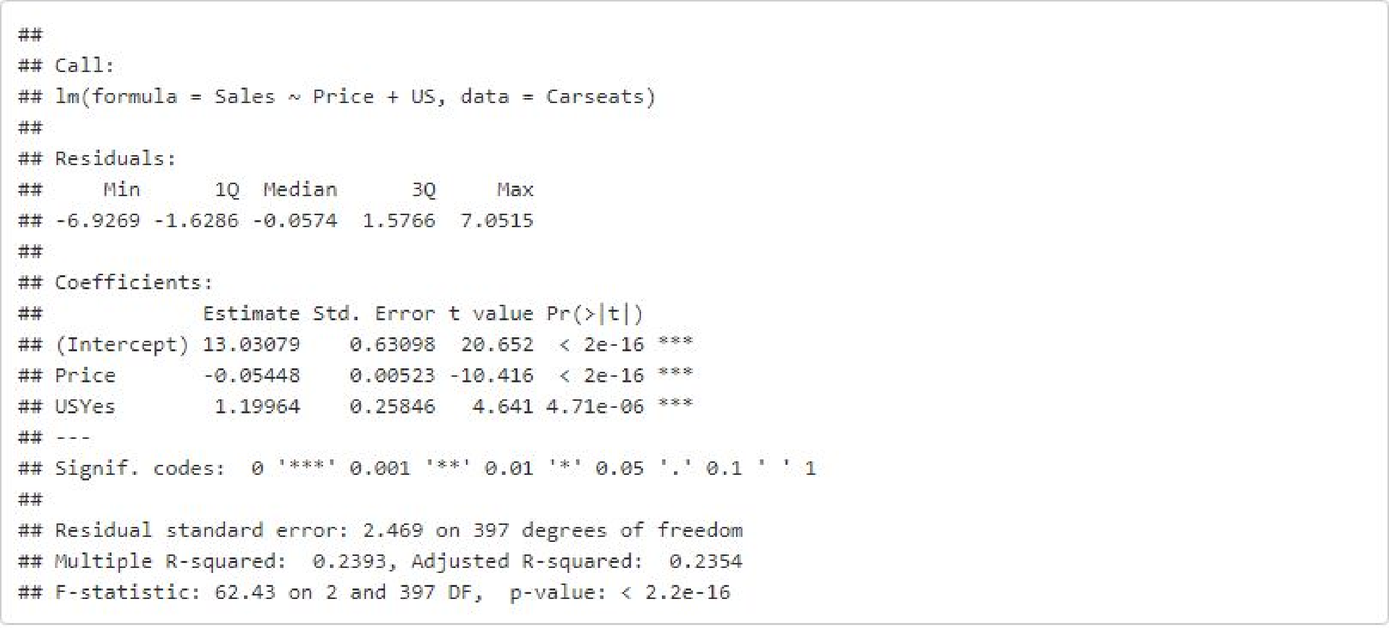
d. For which of the predictors can you reject the null hypothesis H0 : βj = 0?

The Price and US predictors are the only two that can reject the null hypothesis H0 : βj = 0.

e. On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.



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

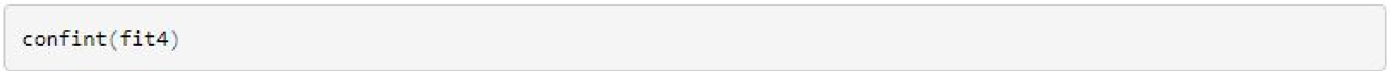


```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

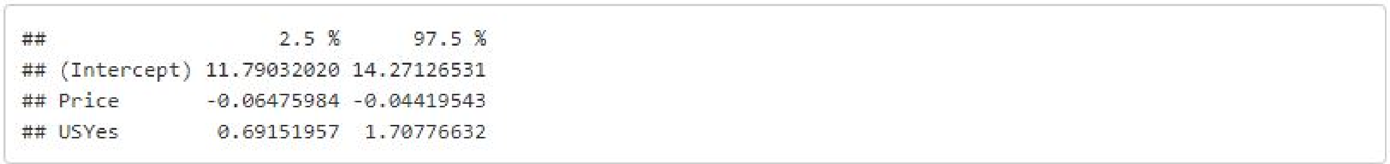
f. How well do the models in (a) and (e) fit the data?

The models in (a) and (e) fit the data well. More specifically, the r squared values in each model are both 0.2393 which indicates that 23.93% of variability exists. Even though a low r squared value does not indicate a good model it also does not indicate a bad model, but the higher the r squared value the better. Additionally, both models can be identified as good models due to both of them producing a significant p-value of 2.2e-16.

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

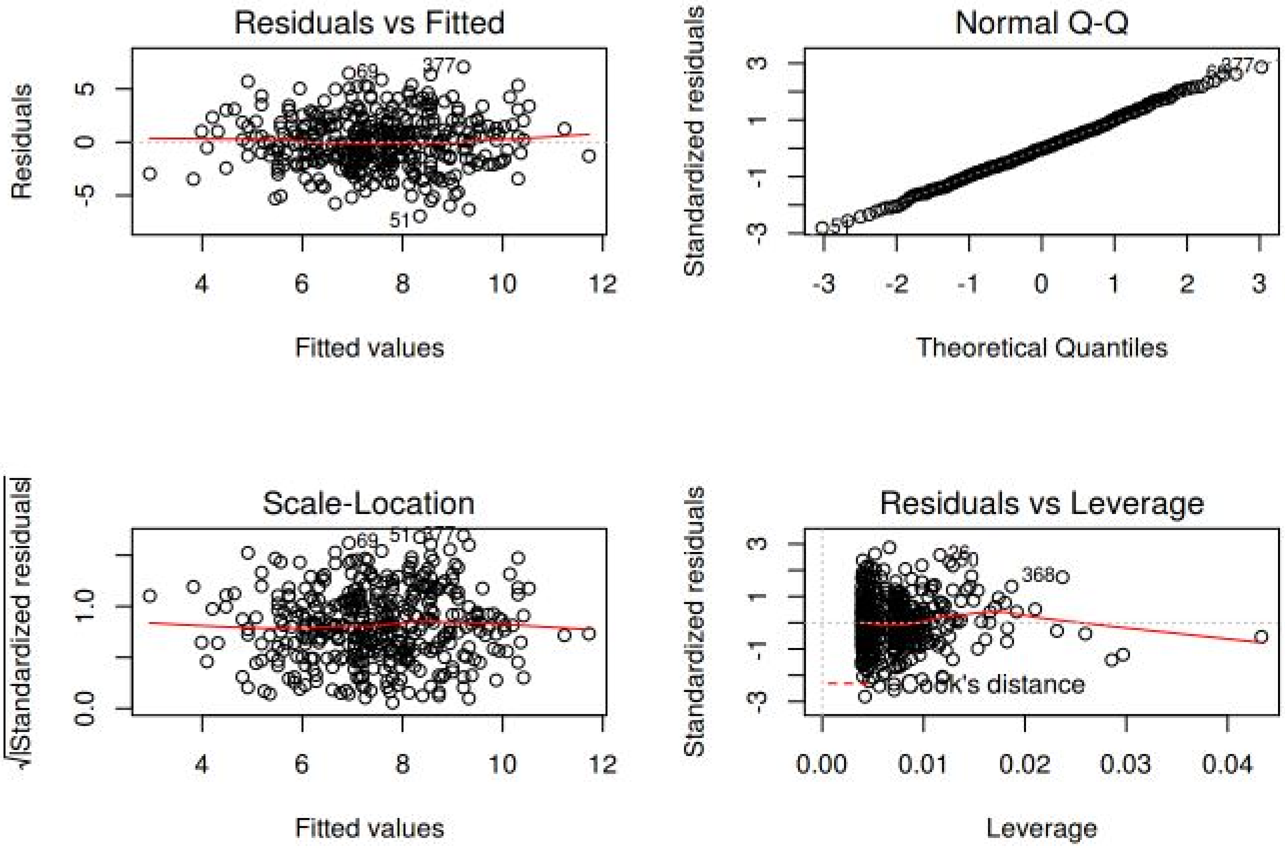


```
confint(fit4)
```



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

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



From the Residuals vs Leverage plot it can be concluded that a small portion of outliers exist as well as various leverage points that are greater than (p+1)/n(0.01).

Question-12

This problem involves simple linear regression without an intercept.

a. Recall that the coefficient estimate β̂ for the linear regression of Y onto X without an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X?

The only circumstance where the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X is

iff $\sum_j x_j^2 = \sum_j y_j^2$

b. Generate an example in R with n = 100 observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.

```r
set.seed(1)
x <- 1:100
sum(x^2)
```

```
## [1] 338350
```

```r
y <- 2 * x + rnorm(100, sd = 0.1)
sum(y^2)
```

```
## [1] 1353606
```

```r
fit.Y <- lm(y ~ x + 0)
fit.X <- lm(x ~ y + 0)
summary(fit.Y)
```

```
## 
## Call:
## lm(formula = y ~ x + 0)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.223590 -0.062560  0.004426  0.058507  0.230926 
## 
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)    
## x 2.0001514  0.0001548   12920   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.09005 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1 
## F-statistic: 1.669e+08 on 1 and 99 DF,  p-value: < 2.2e-16
```

```r
summary(fit.X)
```

```
## 
## Call:
## lm(formula = x ~ y + 0)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.115418 -0.029231 -0.002186  0.031322  0.111795 
## 
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)    
## y 5.00e-01   3.87e-05   12920   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.04502 on 99 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1 
## F-statistic: 1.669e+08 on 1 and 99 DF,  p-value: < 2.2e-16
```

c. Generate an example in R with n = 100 observations in which the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X.

```
x <- 1:100
sum(x^2)
```

```
## [1] 338350
```

```
y <- 100:1
sum(y^2)
```

```
## [1] 338350
```

```
fit.Y <- lm(y ~ x + 0)
fit.X <- lm(x ~ y + 0)
summary(fit.Y)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49.75 -12.44  24.87  62.18  99.49
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   0.5075     0.0866    5.86 6.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.37 on 99 degrees of freedom
## Multiple R-squared:  0.2575, Adjusted R-squared:   0.25
## F-statistic: 34.34 on 1 and 99 DF,  p-value: 6.094e-08
```

```
summary(fit.X)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49.75 -12.44  24.87  62.18  99.49
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y   0.5075     0.0866    5.86 6.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.37 on 99 degrees of freedom
## Multiple R-squared:  0.2575, Adjusted R-squared:   0.25
## F-statistic: 34.34 on 1 and 99 DF,  p-value: 6.094e-08
```