Assignment 1

Brian Morales

August 31, 2022

We will be using the `Carseats` data set

Part A

```
library(readr)
Carseats <- read_csv("~/Desktop/Fall-2022/Stats-Learning/ALL-CSV-FILES/Carseats.csv", show_col_types = F

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

```
##
## 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
```

Part B

Each coefficient in the model refers to a β_p, the quantified relation between each predictor and response variable, in our linear model $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$. - In our case $\beta_0 = 13.043$ is the average value of $\mathbf{Y}$, `Sales`, when $\mathbf{X}$ is zero or when a company spend $0 on `Price`, not `UrbanYes` and not `USYes` .

- `Price` is $\beta_1 = -0.054$ and this is the average change in $\mathbf{Y}$ associated with a 1-unit increase in the value x_j. In other words, for 100 dollar increase (not sure if its 100's, 1000's) in `Price` the company can expect to sell, $\beta_1 \times 100 = -0.054 \times 100 = -5.4$ less Car Seats sales, on average. There is an extremely low p-value indicating that `Price` and `Sales` have a relation.

- β_2 =`Urban` and β_3 =`US` and these are qualitative parameters. Therefore when `UrbanYes` is true and the parameters `Price`= 0 and `USYes` is false, we will see $\beta_2 \times 100 = -0.02 \times 100 = -2$ decrease in car seats sales, on average (assuming units are in the 100's). Also, given the high p-value in the model this is suggesting there is no relationship between `Urban` and `Sales`
- Similarly with `US`, we will see a 120 increase in car seats sales. On the other hand, `USYes` has significantly low p-value therefore showing evidence of some relation with `Sales`.

Part C

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3} = 13.043 - 0.054\text{Price} - 0.02\text{UrbanYes} + 1.2\text{USYes}$$

Each $\hat{\beta}_p$ corresponds to the coefficient printed out in the summary table. Each $x_{i,p}$ corresponds to its predictor according to its coefficient.

Part D

Looking at the summary table below, if our significance level is $\alpha = 0.05$ we can reject the null hypothesis H_0. For `Price` and `USYes` the p-value< 0.05 implicating that `Price` and `USYes` predictors has an association to the response, `Sales`. `UrbanYes` has a p-value> 0.05, therefore we can't reject the null hypothesis proffering that their is no relation between `UrbanYes` and `Sales`.

```
summary(carsts_lm)
```

```
##
## 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
```

Part E

```r
carsts_lms = lm(Sales ~ Price + US, Carseats)
summary(carsts_lms)
```

```
##
## 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
```

Above we have a summary of the reduced model - excluding `UrbanYes`. Notice that all the p-values related to each coefficient is significant and the F-statistics has increased.

Part F

Observing P-value of the `carsts_lms`(Part E) model, the p-value is $2.2e^{-16} < 0.05$ so we reject the model. This signifies that we need one of the predictors. However if we look at `carsts_lm`(Part A) the p-value is also < 0.05 hinting that we need more predictors. Lets run an a full F-test:

```r
anova(carsts_lms, carsts_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + US
## Model 2: Sales ~ Price + Urban + US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    397 2420.9
## 2    396 2420.8  1   0.03979 0.0065 0.9357
```

Notice that the p-values associated with the F-test is large, demonstrating that `carsts_lms` is sufficient. Nevertheless, the RSS and the adjusted R^2 in both models are very similar. What I conclude is that both models fits the data fairly equally.The `carsts_lms` fits the model somewhat better if you want to be precise.

Part G

```r
#seB = summary(carsts_lms)$coefficients[2, 2]
#beta1 = coefficients(carsts_lms)[2]
#n = nrow(Carseats)
#CI = c( beta1 - qt(0.975, df=n-3)*seB, beta1 + qt(0.975, df=n-3)*seB)
#CI

confint(carsts_lms)
```

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

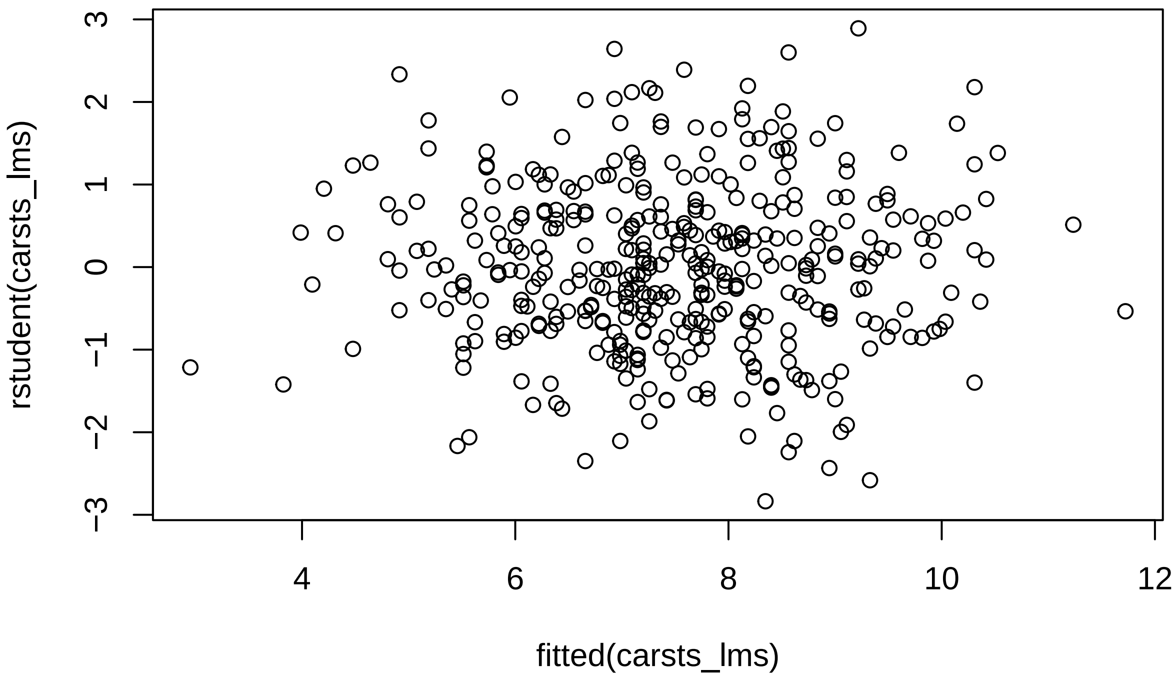
Here we are 95% confident that the true value lies around $(-0.064, -0.04)$ for `Price` and $(0.69, 1.70)$ for `USYes`.

Part H

```r
library(car)
```

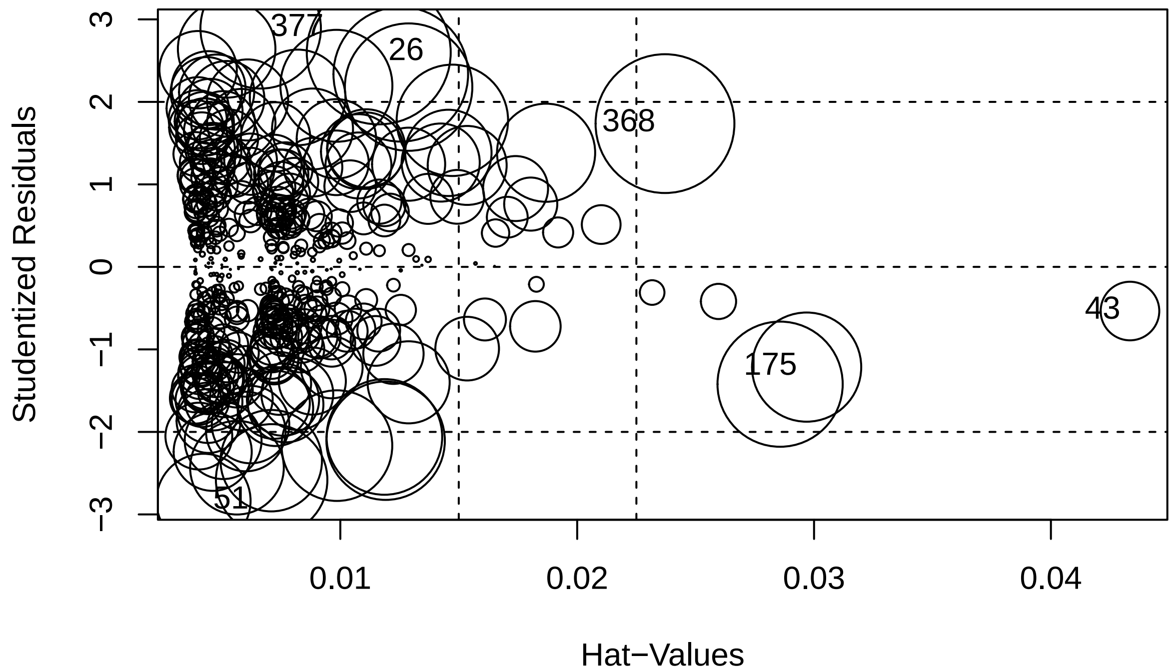
```
## Loading required package: carData
```

```r
#cooksd = cooks.distance(carsts_lms)
#plot(cooksd, pch = "*", cex = 2, main="Influential Obs by Cooks Distance")
plot(fitted(carsts_lms), rstudent(carsts_lms))
```



Looking at the studentized Residual vs Fitted there looks to be no potential outliers. Every points stay in the same range of $[-3, 3]$, which is what we typically look for.

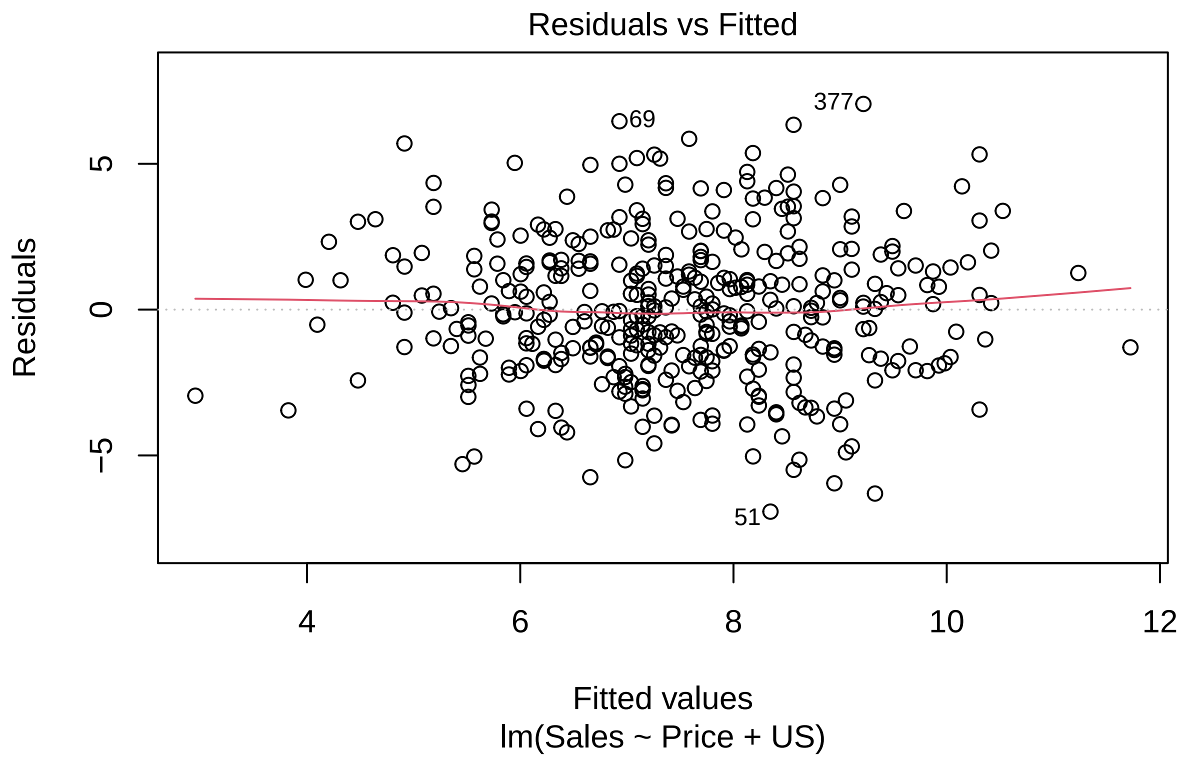
```
influencePlot(carsts_lms)
```



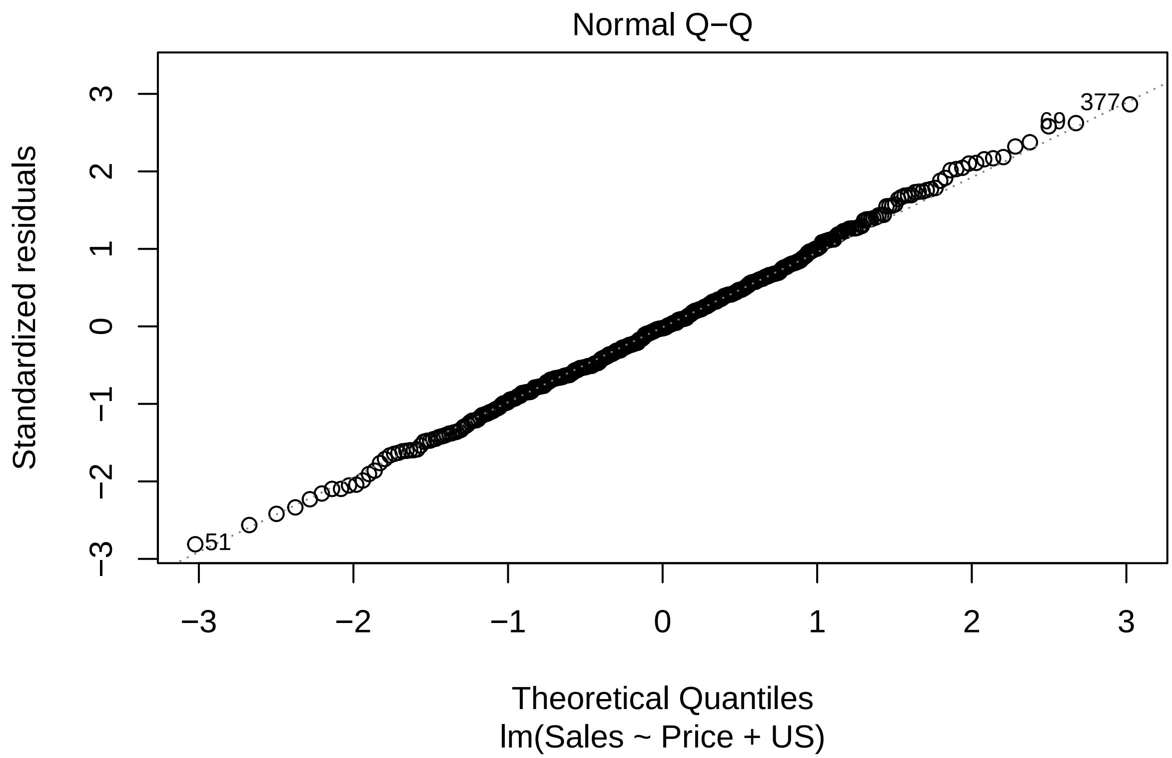

```
##         StudRes         Hat       CookD
## 26   2.5996518 0.011621599 0.026109457
## 43  -0.5349931 0.043337657 0.004329756
## 51  -2.8358431 0.004224147 0.011173381
## 175 -1.2144859 0.029686718 0.015024314
## 368  1.7366086 0.023707048 0.024287363
## 377  2.8915213 0.006637175 0.018282191
```

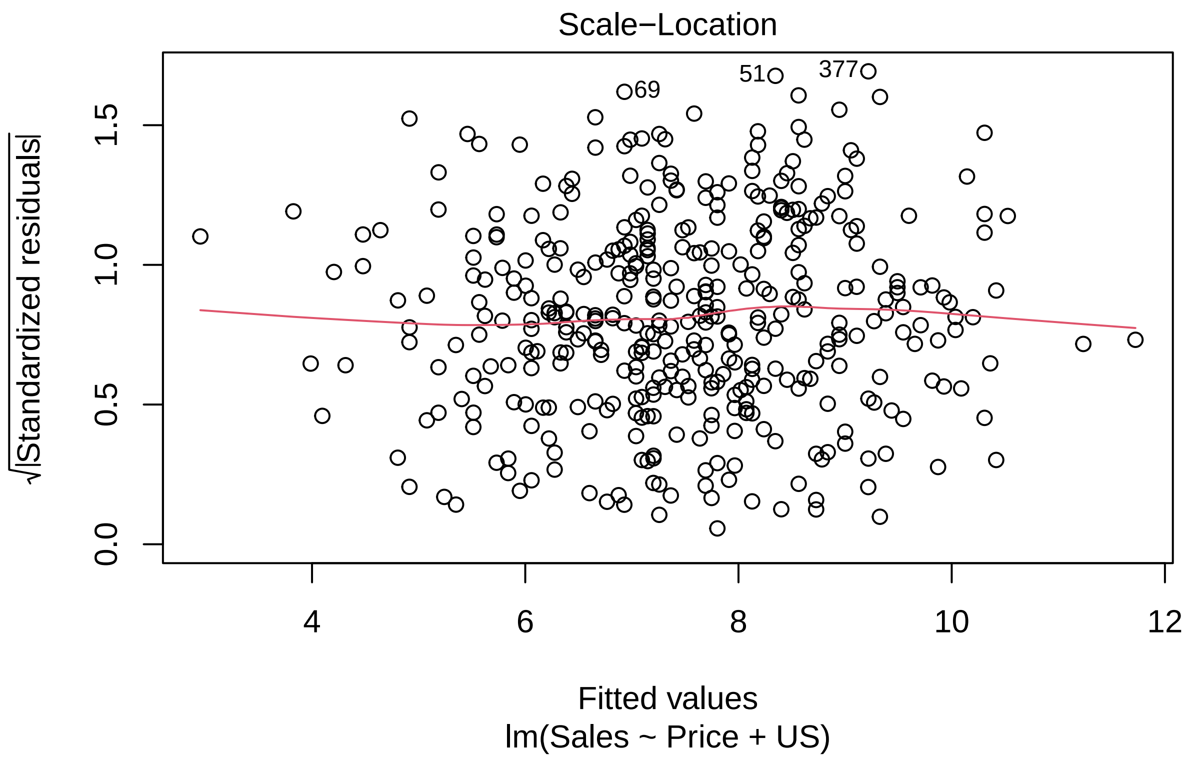
Observing the Influence Plot, there are a few observations that can be removed such as point 26, 377, and 210.

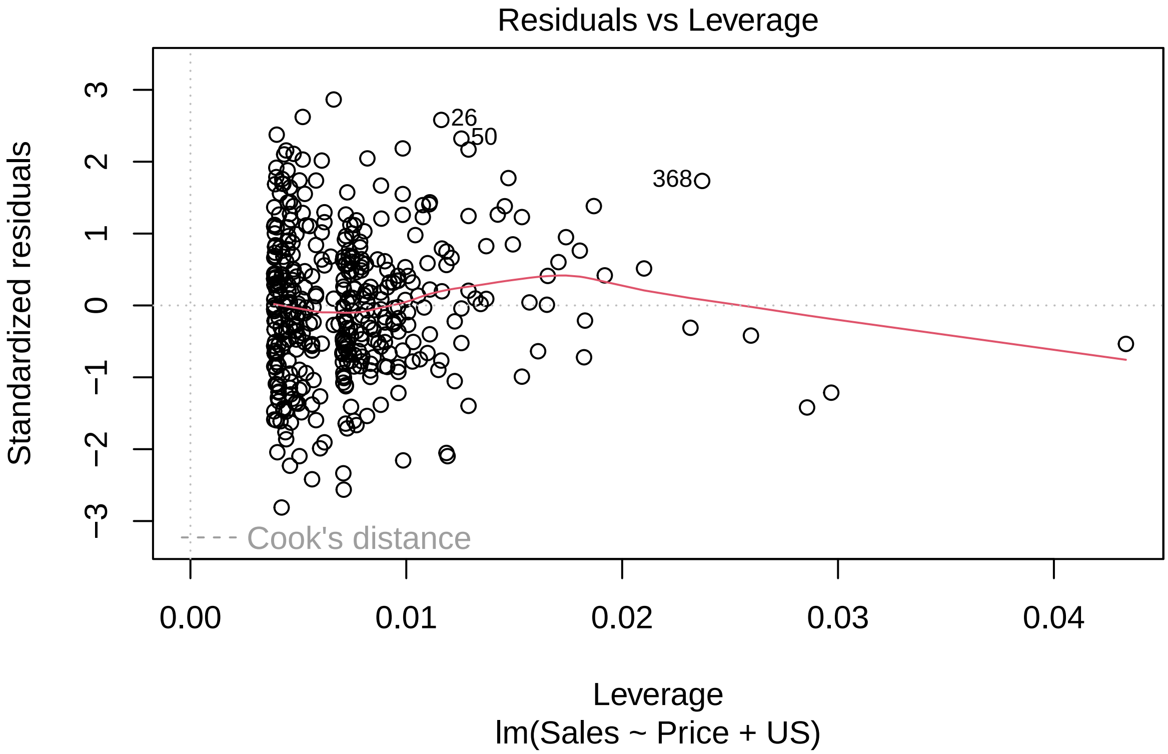
```
plot(carsts_lms)
```



Residuals vs Fitted
Residuals
Fitted values
lm(Sales ~ Price + US)
377
69
51
5
0
−5
4
6
8
10
12


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ Price + US)
377
69
51
−3
−2
−1
0
1
2
3


Scale−Location
69
51
377
√|Standardized residuals|
0.0
0.5
1.0
1.5
4
6
8
10
12
Fitted values
lm(Sales ~ Price + US)



Observing the Residuals vs Leverage, there are a few observations the are greater than the average leverage for all observations, $\frac{p+1}{n} = \frac{3}{400} = 0.0075$. This represents that some observations have high leverage.