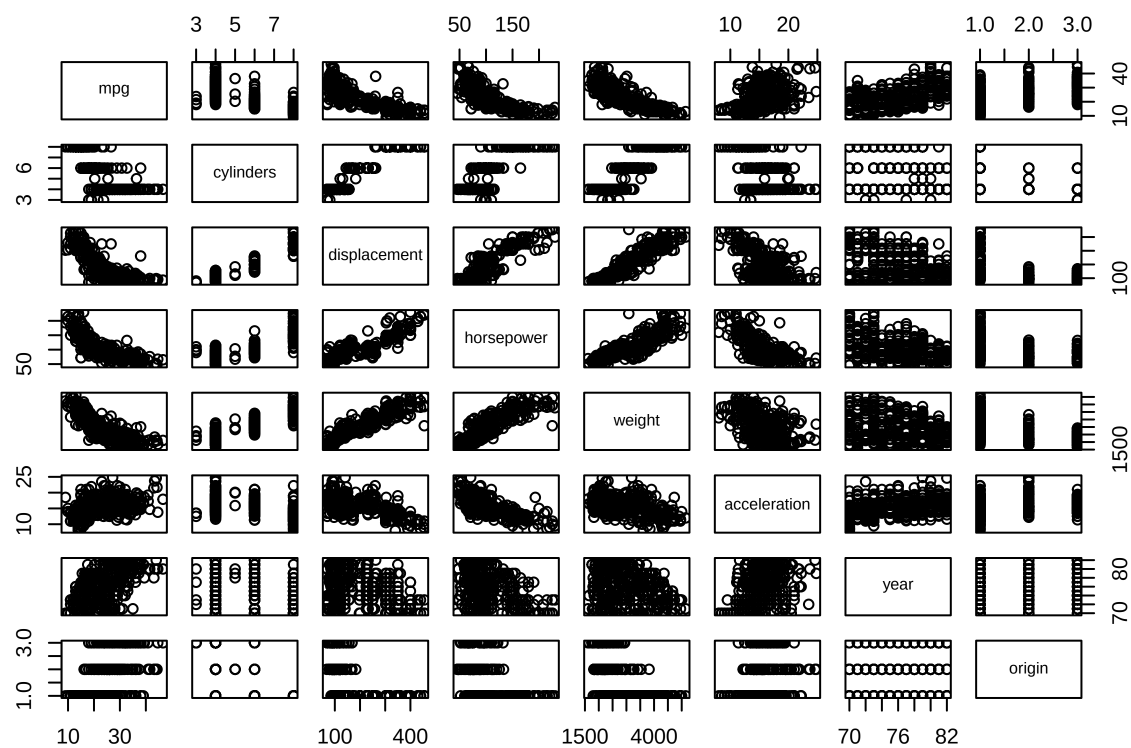
Lab02 Brandon Habschied

```r
Auto <- read.csv("C:/Users/brand/OneDrive/Desktop/School/DataDrivenDiscovery/Labs/Lab2/Auto.csv")
View(Auto)
Auto <- na.omit(Auto)
# str(Auto)
# convert horsepower from char to double so it can be in the scatterplot
Auto$horsepower <- as.numeric(Auto$horsepower)
```

```
## Warning: NAs introduced by coercion
```

```r
# scatterplot the numerical variables only
Auto_num <- Auto[, -ncol(Auto)]
pairs(Auto_num)
```




```r
round(cor(Auto_num), digits = 3)
```

```
##                    mpg cylinders displacement horsepower weight acceleration
```

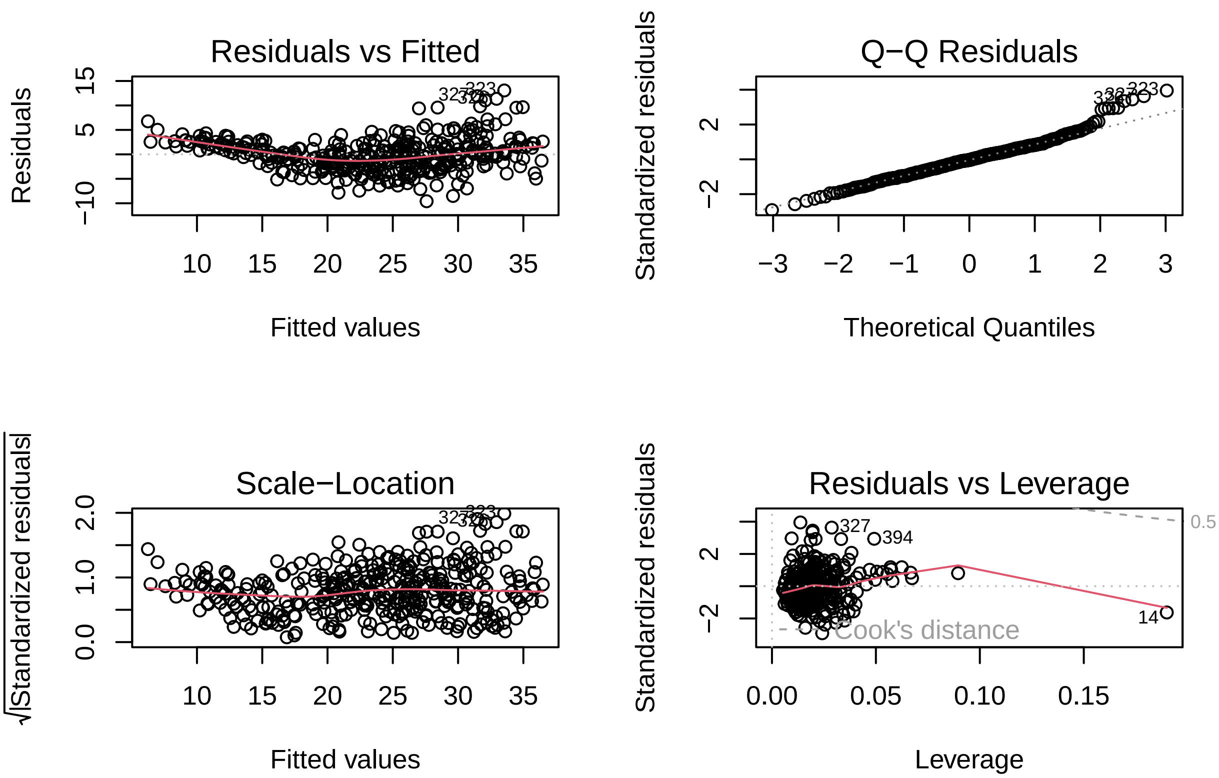
```
## mpg           1.000    -0.776       -0.804         NA -0.832        0.422
## cylinders    -0.776     1.000        0.951         NA  0.897       -0.504
## displacement -0.804     0.951        1.000         NA  0.933       -0.544
## horsepower       NA        NA           NA          1     NA           NA
## weight       -0.832     0.897        0.933         NA  1.000       -0.420
## acceleration  0.422    -0.504       -0.544         NA -0.420        1.000
## year          0.581    -0.347       -0.370         NA -0.308        0.283
## origin        0.564    -0.565       -0.611         NA -0.581        0.210
##                year origin
## mpg           0.581  0.564
## cylinders    -0.347 -0.565
## displacement -0.370 -0.611
## horsepower       NA     NA
## weight       -0.308 -0.581
## acceleration  0.283  0.210
## year          1.000  0.184
## origin        0.184  1.000
```

```r
linReg <- lm(mpg ~ . , data = Auto_num)
summary(linReg)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = Auto_num)
##
## 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
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```r
#
# 1. There is a relationship between the predictors and the response as noted by
# the large F statistic of 252.4 and the extremely small p-value of 2.2e-16
#
# 2. The predictors that appear to have significantly significant relationships
# are the ones with the extremely small p-values:
```

```r
# weight, year, origin, displacement
#
# 3.The coefficient for year is 0.75 which suggests that each year, newer
# vehicles' mpg will increase by .75 should all other variables stay the same.
#
par(mfrow=c(2,2))
plot(linReg)
```

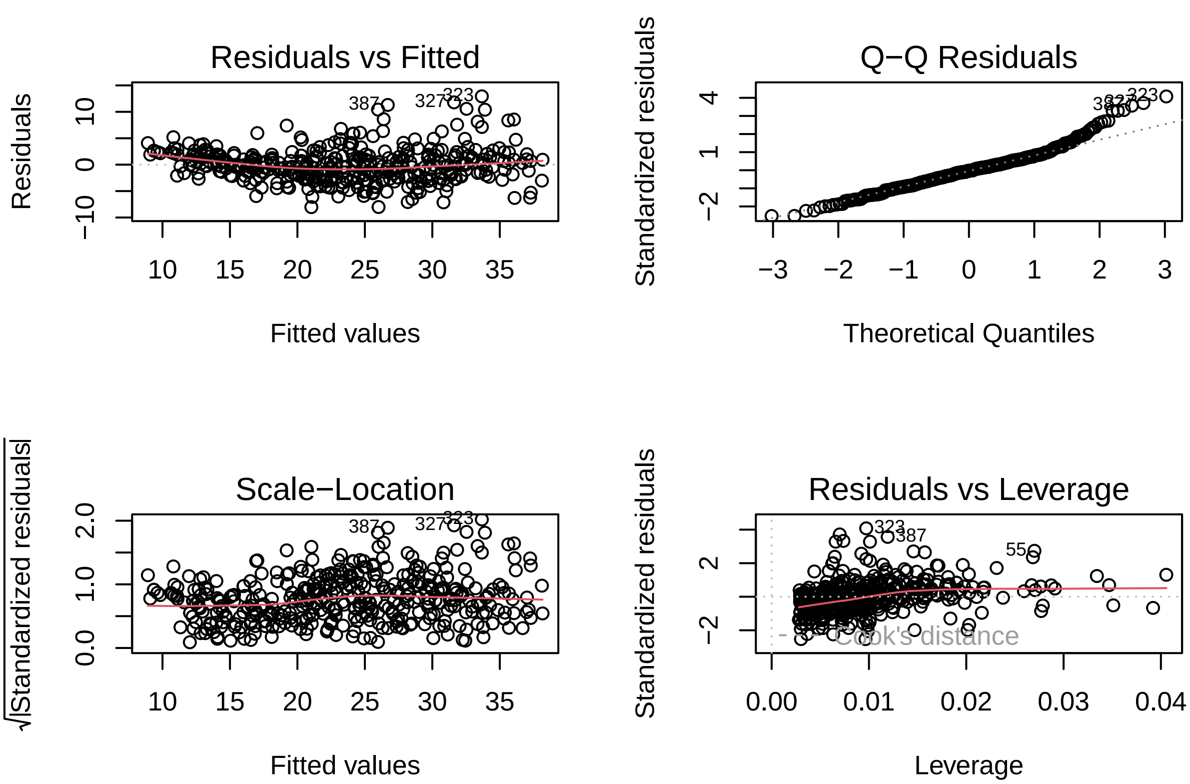



```r
# There are a few unusually large outliers found on the residual plots,
# 327,323, 336 stand out immediately. The QQ Residuals also has a cluster of
# unusually large outliers. The leverage plot has a specific instance 14 with
# an abnormally large leverage.
linReg2 <- lm(mpg ~ weight * year, data = Auto_num)
summary(linReg2)
```

```
##
## Call:
## lm(formula = mpg ~ weight * year, data = Auto_num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0341 -1.9851 -0.0912  1.6987 12.9292
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -1.124e+02  1.280e+01  -8.781  < 2e-16 ***
## weight       2.821e-02  4.376e-03   6.447 3.34e-10 ***
## year         2.068e+00  1.699e-01  12.171  < 2e-16 ***
## weight:year -4.672e-04  5.857e-05  -7.977 1.66e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.187 on 393 degrees of freedom
## Multiple R-squared:  0.8354, Adjusted R-squared:  0.8341
## F-statistic: 664.9 on 3 and 393 DF,  p-value: < 2.2e-16
```

```
plot(linReg2)
```

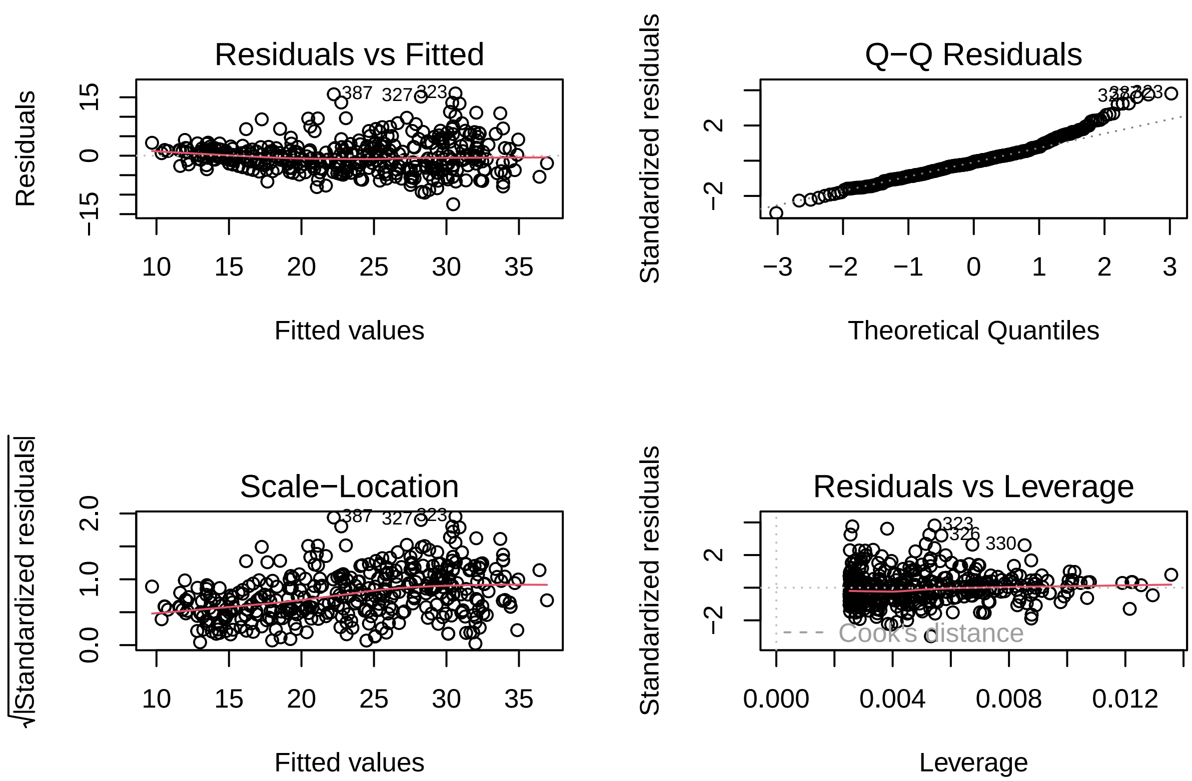



```
linReg3 <- lm(mpg ~ log(weight), data = Auto_num)
summary(linReg3)
```

```
##
## Call:
## lm(formula = mpg ~ log(weight), data = Auto_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4658  -2.6579  -0.2947   1.9395  15.9787
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 210.5391     5.9837   35.19   <2e-16 ***
## log(weight) -23.5050     0.7516  -31.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.203 on 395 degrees of freedom
## Multiple R-squared:  0.7123, Adjusted R-squared:  0.7116
## F-statistic: 978.1 on 1 and 395 DF,  p-value: < 2.2e-16
```

```r
plot(linReg3)
```

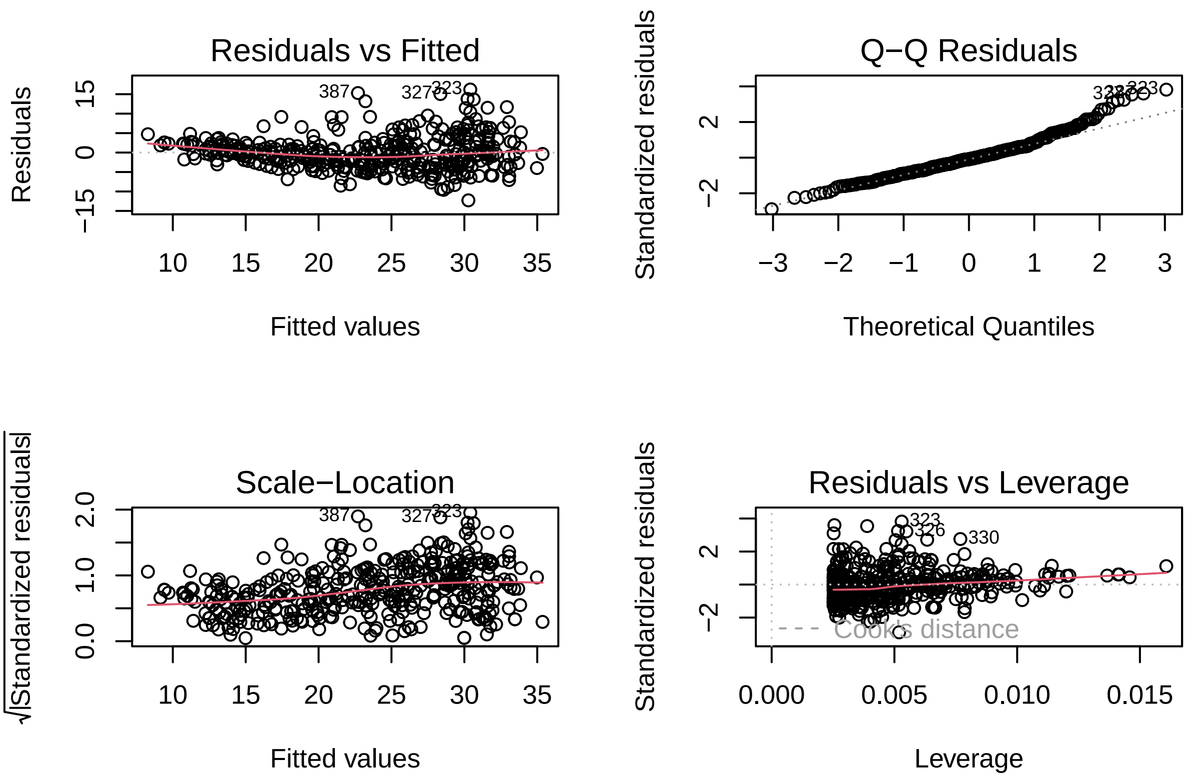



```r
linReg4 <- lm(mpg ~ sqrt(weight), data = Auto_num)
summary(linReg4)
```

```
##
## Call:
## lm(formula = mpg ~ sqrt(weight), data = Auto_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2769  -2.8948  -0.3705   2.0839  16.1925
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  69.84709    1.52239   45.88    <2e-16 ***
## sqrt(weight) -0.85860    0.02793  -30.74    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 395 degrees of freedom
## Multiple R-squared:  0.7052, Adjusted R-squared:  0.7044
## F-statistic: 944.8 on 1 and 395 DF,  p-value: < 2.2e-16
```

```r
plot(linReg4)
```




```r
par(mfrow=c(1,1))
# The interactions I tested above all seem to be statistically significant as
# they all have very small p-values and large F statistics of
# 650, 967, and 935. Using the different interactions caused the leverage
# line of best fit to be more aligned with cook's distance.

# PART 2
library(ISLR)
```

```
##
## Attaching package: 'ISLR'

## The following object is masked _by_ '.GlobalEnv':
##
##     Auto
```

```r
data("Carseats")
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 ...
```

```r
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

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

```
## 
## 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
```

```r
# The coefficient for price being -0.054 suggests that for every increase in
# in price by 1, sales will drop by 0.054. The coefficient for Urban of
# -0.0219 suggests that a carseat in an urban setting (1) will have 0.0219 less
# sales than one in a non-urban setting (0). The coefficient for US suggests that a
# carseat made in the US (1) is likely to have 1.2 more sale units than one not
# made in the US (0).
#
# C) Sales = 13.043 + (-0.054 * Price) + (-0.219 * Urban) + (1.2 * US) + error
#
# D) We can reject the null hypothesis for the predictors Price and US due to
# their extremely low p-values
# e)
mReg2 <- lm(Sales ~ Price + US, data = Carseats)
summary(mReg2)
```

```
##
## 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
```

```r
# f) The models above have RSE values that are close to 0, which indicates a
# good fit for the data. Additionally, they both have relatively low F stats
# which indicates that the model is statistically significant.
#
# g) using mReg2 from e) we can determine a 95% confidence interval using 1.96
# as our critical value.

cat("Sales 95% CI: [", 13.03079 - 1.96 * 0.63098, ",", 13.03079 + 1.96 * 0.63098, "]")
```

```
## Sales 95% CI: [ 11.79407 , 14.26751 ]
```

```r
cat("Price 95% CI: [", -0.05448 - 1.96 * 0.00523, ",", -0.05448 + 1.96 * 0.00523, "]")
```

```
## Price 95% CI: [ -0.0647308 , -0.0442292 ]
```

```r
cat("US 95% CI: [", 1.19964 - 1.96 * 0.25846, ",", 1.19964 + 1.96 * 0.25846, "]")
```

```
## US 95% CI: [ 0.6930584 , 1.706222 ]
```