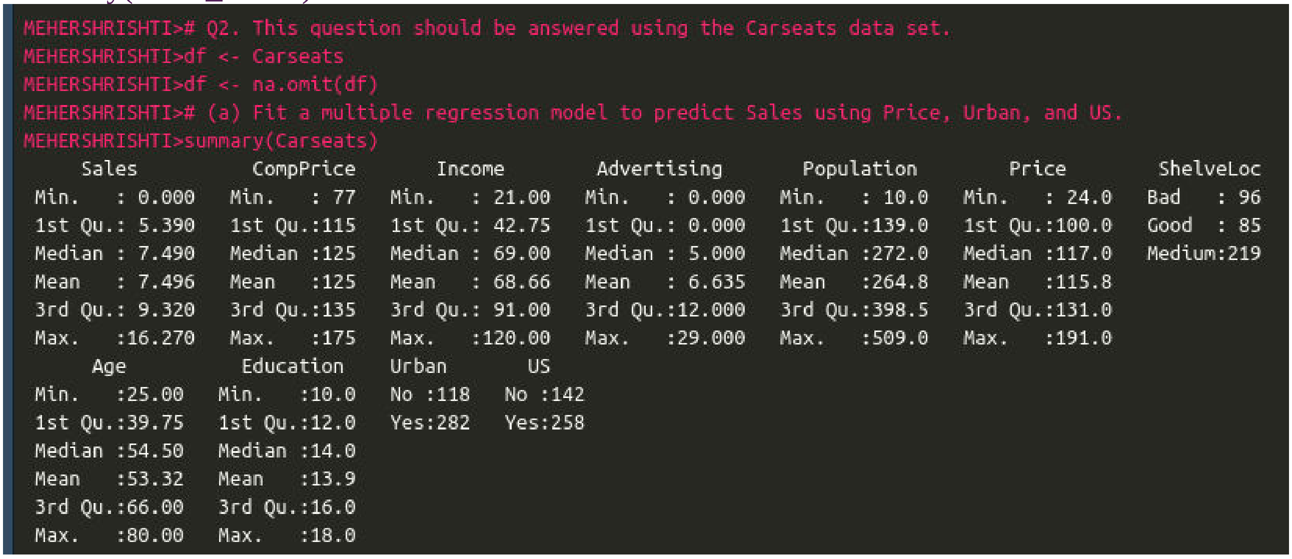
# Meher Shrishti Nigam
# 20BRS1193

# EDA LAB – 2
# 6 / 1 / 23

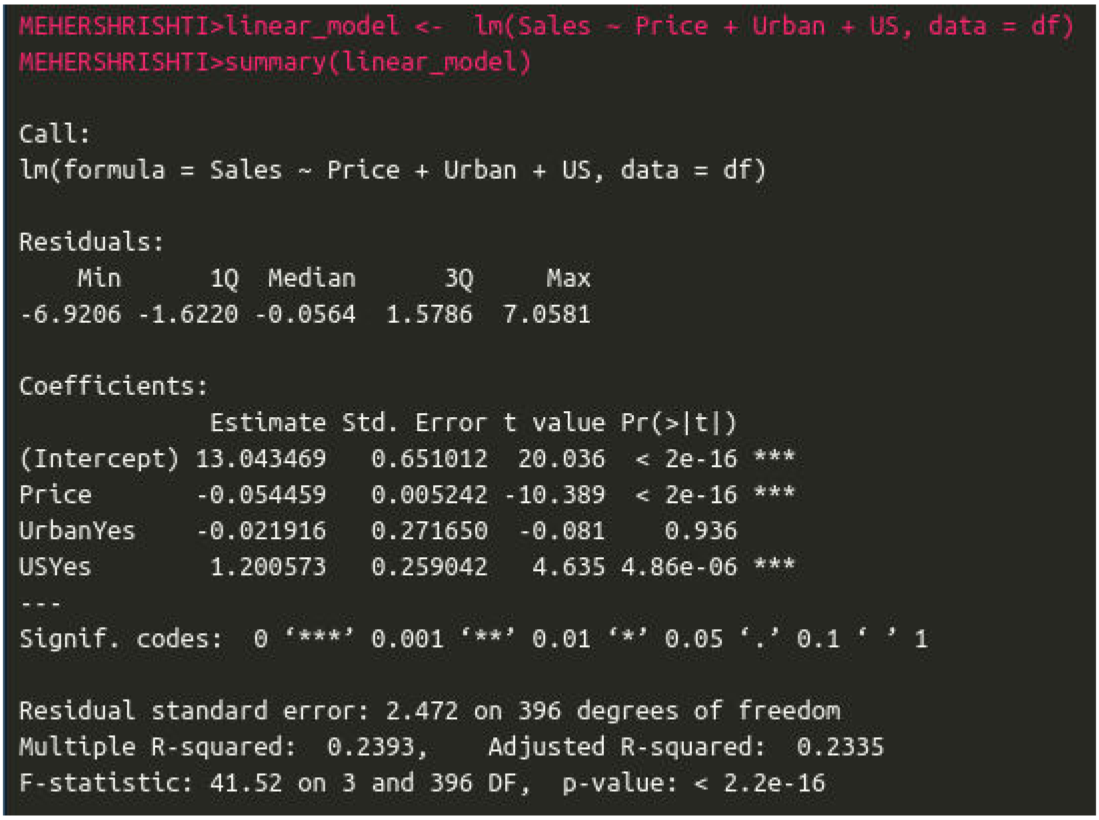
# Meher Shrishti Nigam
# 20BRS1193
# EDA Lab 2
options(prompt="MEHERSHRISHTI>", continue =" ")
# options(prompt=">", continue =" ")
# EDA-LAB-EXPERIMENT-2 (Date-6/1/2023)
library(ISLR)

**# Q2. This question should be answered using the Carseats data set.**
df <- Carseats
df <- na.omit(df)

**# (a) Fit a multiple regression model to predict Sales using Price, Urban, and US.**
summary(Carseats)
linear_model <-  lm(Sales ~ Price + Urban + US, data = df)
summary(linear_model)



```
MEHERSHRISHTI># Q2. This question should be answered using the Carseats data set.
MEHERSHRISHTI>df <- Carseats
MEHERSHRISHTI>df <- na.omit(df)
MEHERSHRISHTI># (a) Fit a multiple regression model to predict Sales using Price, Urban, and US.
MEHERSHRISHTI>summary(Carseats)
     Sales          CompPrice       Income        Advertising       Population        Price        ShelveLoc  
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0   Min.   : 24.0   Bad   : 96  
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0   Good  : 85  
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0   Median :117.0   Medium:219  
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8   Mean   :115.8               
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0               
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0   Max.   :191.0               
      Age          Education    Urban       US     
 Min.   :25.00   Min.   :10.0   No :118   No :142  
 1st Qu.:39.75   1st Qu.:12.0   Yes:282   Yes:258  
 Median :54.50   Median :14.0                      
 Mean   :53.32   Mean   :13.9                      
 3rd Qu.:66.00   3rd Qu.:16.0                      
 Max.   :80.00   Max.   :18.0                      
```



```
MEHERSHRISHTI>linear_model <-  lm(Sales ~ Price + Urban + US, data = df)
MEHERSHRISHTI>summary(linear_model)

Call:
lm(formula = Sales ~ Price + Urban + US, data = df)

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!**

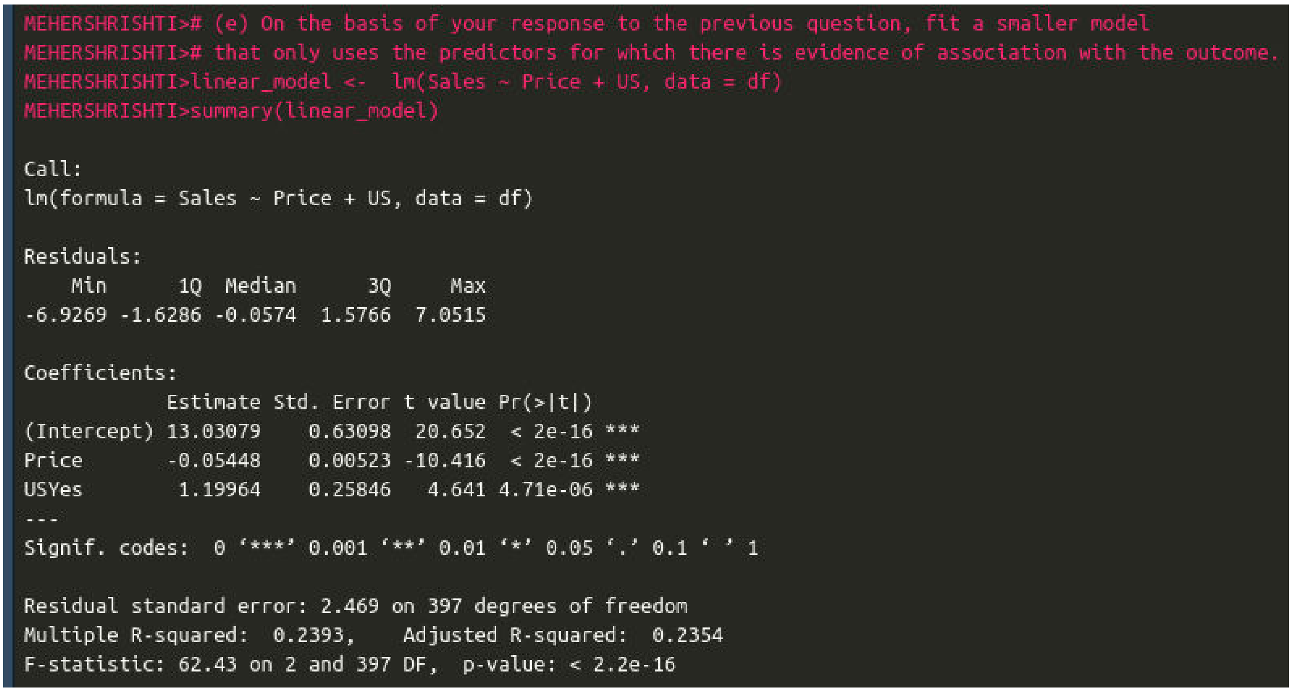
USYes: Our linear model predicts that USYes and Sales are correlated. The coefficient is positive, thus, when USYes increases, Sales decreases.
In the Price attribute, we observe a low p value of the t statistic. Thus, our linear model predicts that Price and Sales are correlated.
The coefficient is negative, thus, when Price increases, Sales decreases.
We notice that UrbanYes, a variable that gives us information about the location of the store, has no relation with number of sales in our linear model.

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

Sales = 13.043469 -0.054459 Price -0.021916 UrbanYes + 1.200573 USYes
We can remove the UrbanYes function as well.

**# (d) For which of the predictors can you reject the null hypothesis H0 : βj = 0?**
Price and USYes, based on the  p-value of 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.**
linear_model <-  lm(Sales ~ Price + US, data = df)
summary(linear_model)



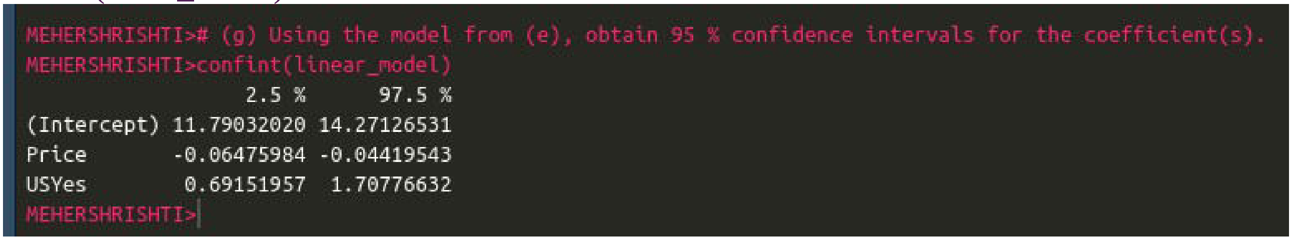
```
MEHERSHRISHTI># (e) On the basis of your response to the previous question, fit a smaller model
MEHERSHRISHTI># that only uses the predictors for which there is evidence of association with the outcome.
MEHERSHRISHTI>linear_model <-  lm(Sales ~ Price + US, data = df)
MEHERSHRISHTI>summary(linear_model)

Call:
lm(formula = Sales ~ Price + US, data = df)

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 of the linear regressions fit the data similarly based on RSE and R2, with linear regression from (e) significantly doing very little better.

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

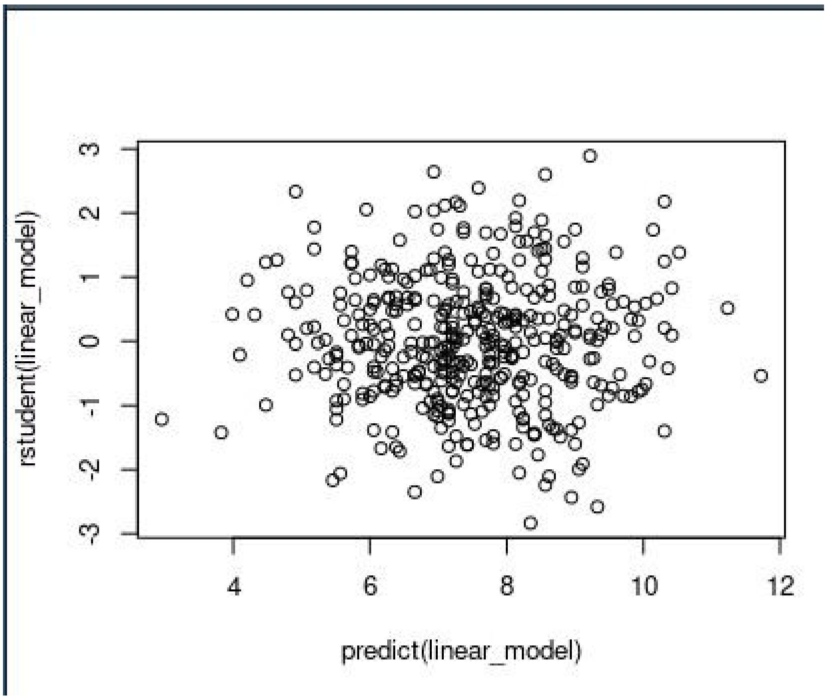
confint(linear_model)



```
MEHERSHRISHTI># (g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).
MEHERSHRISHTI>confint(linear_model)
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
MEHERSHRISHTI>
```

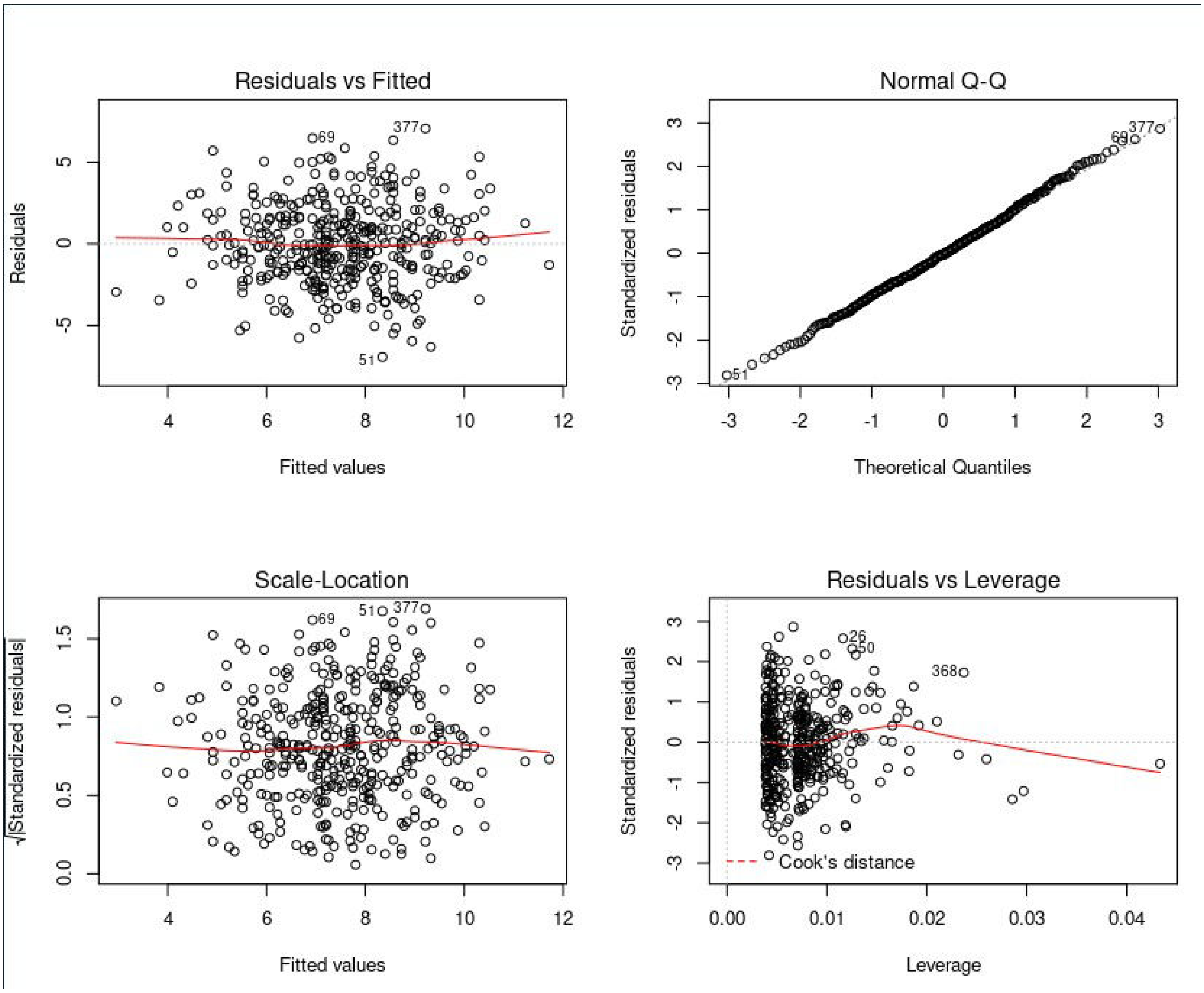
# (h) Is there evidence of outliers or high leverage observations in the model from (e)?

plot(predict(linear_model), rstudent(linear_model))



No probable outliers are inferred from the linear regression because all residuals are within the range of -3 to 3.

```
par(mfrow = c(2, 2))
plot(linear_model)
```


Residuals vs Fitted
Residuals
Fitted values
377
69
51
Normal Q-Q
Standardized residuals
Theoretical Quantiles
377
69
51
Scale-Location
√|Standardized residuals|
Fitted values
51
377
69
Residuals vs Leverage
Standardized residuals
Leverage
26
50
368
Cook's distance