chapter 3

LinearRegression

02/10/2024

#chapter 3 (9)

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.

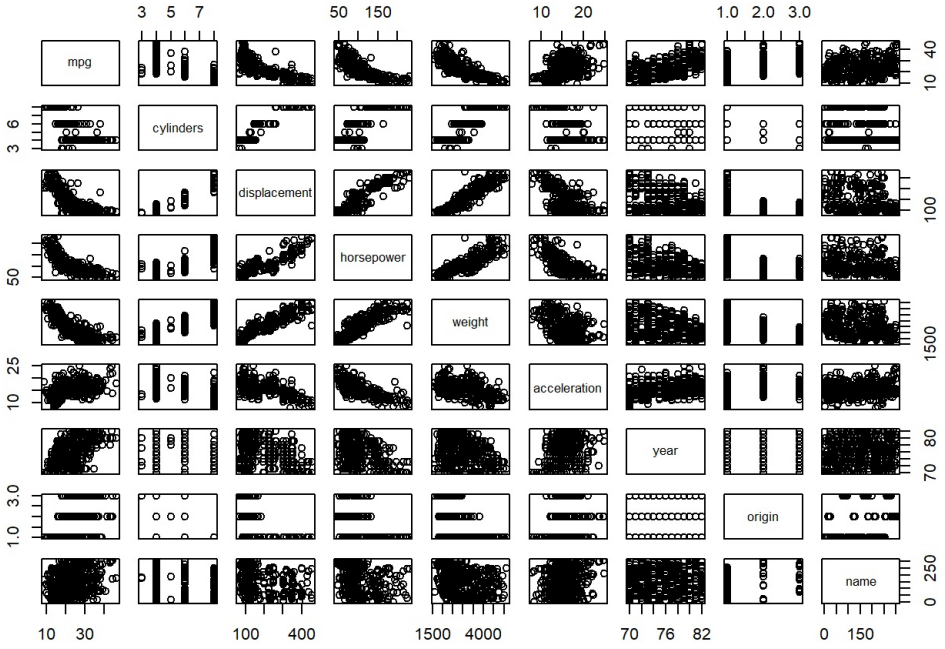
```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.3.3
```

```
?Auto
```

```
## starting httpd help server ... done
```

```
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.

```
cor(Auto[ ,-9])
```

```
##                     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
```

INTERPRETATION #MPG shows a strong negative correlation with factors like cylinders, displacement, horsepower, and weight, indicating that as these characteristics increase, fuel efficiency #MPG has a positive relationship with acceleration, year, and origin, which means that cars that accelerate faster, are newer, and come from certain regions generally have better fuel efficiency. #Finally, cars from certain regions (origin) tend to have higher fuel efficiency, smaller engines, and lighter weight (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: i. Is there a relationship between the predictors and the response? ii. Which predictors appear to have a statistically significant relationship to the response? iii. What does the coefficient for the year variable suggest?

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

```
##
## 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
```

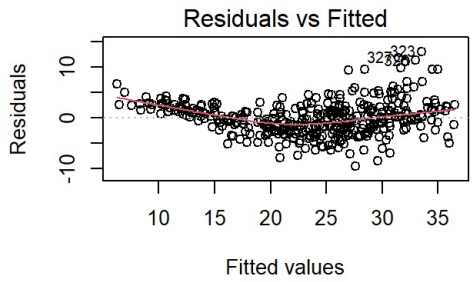
INTERPRETATION #The minimum residual is -9.59, and the maximum is 13.06, showing the range of prediction errors #Cylinders (-0.493): A negative but statistically insignificant effect on mpg (p-value: 0.1278) #A small positive effect on mpg, with a significant p-value (0.0084), meaning that larger engine displacement slightly increases mpg. #Displacement is also significant but has a smaller effect, while other variables like cylinders, horsepower, and acceleration are not statistically significant the coefficient for the year variable

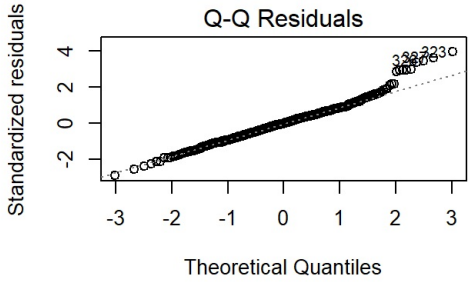
```
coef(mpg)[7]
```

```
##      year
## 0.7507727
```

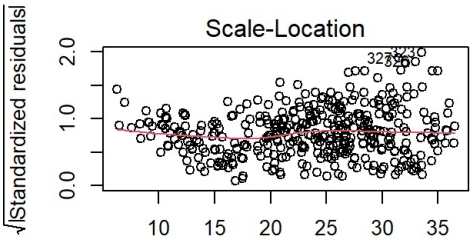
d. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually highleverage?

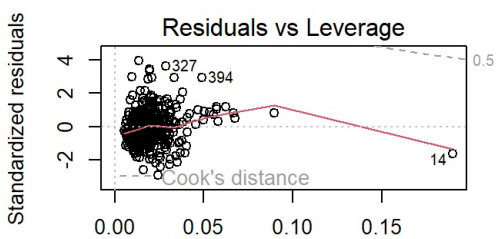
```
par(mfrow=c(2,2))
plot(mpg)
```





INTERPRETATION #Normal Q-Q





residuals should fall along the diagonal line. If they deviate significantly from this line (especially in the tails), it indicates that the residuals are not normally distributed #A scale-location plot with random scatter suggests that the residual variance is constant. #If no extreme points are identified in the residuals vs leverage plot, the model is not overly influenced by outliers

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

```
summary(lm(formula = mpg ~ . * ., data = Auto[, -9]))
```

```
##
## Call:
## lm(formula = mpg ~ . * ., data = Auto[, -9])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.548e+01  5.314e+01   0.668  0.50475
## cylinders                   6.989e+00  8.248e+00   0.847  0.39738
## displacement               -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                  5.034e-01  3.470e-01   1.451  0.14769
## weight                      4.133e-03  1.759e-02   0.235  0.81442
## acceleration               -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                        6.974e-01  6.097e-01   1.144  0.25340
## origin                     -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders:displacement     -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower        1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight            3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration      2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year             -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin            4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower    -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight         2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year           5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin         2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight          -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration    -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year            -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin           2.233e-03  2.930e-02   0.076  0.93931
## weight:acceleration         2.346e-04  2.289e-04   1.025  0.30596
## weight:year                -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin              -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year           5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin         4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                 1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

INTERPRETATION #Weight, horsepower, cylinders, and some interactions do not have a significant direct impact on mpg when considered alongside other variables #Displacement (engine size) has a significant negative effect, suggesting larger engines reduce mpg #Weight has an almost negligible and statistically insignificant impact on mpg #Origin has a significant negative effect, meaning cars from certain regions tend to have lower fuel efficiency (f) Try a few different transformations of the variables, such as log(X), √ X, X2. Comment on your findings.

the standard linear model, we get an adjusted R2

```
mpg2 <- lm(mpg ~ . - name + I(weight^2) + I(displacement^2) + I(horsepower^2) + I(year^2), data = Auto)
summary(mpg2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name + I(weight^2) + I(displacement^2) +
##     I(horsepower^2) + I(year^2), data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8607 -1.4222  0.0293  1.3596 11.7816
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.121e+02  6.969e+01   5.912 7.51e-09 ***
## cylinders          5.892e-01  3.154e-01   1.868 0.062514 .
## displacement      -4.415e-02  1.929e-02  -2.289 0.022606 *
## horsepower        -1.861e-01  3.928e-02  -4.737 3.06e-06 ***
## weight            -9.982e-03  2.485e-03  -4.016 7.13e-05 ***
## acceleration      -1.837e-01  9.631e-02  -1.907 0.057279 .
## year              -1.003e+01  1.838e+00  -5.455 8.86e-08 ***
## origin             5.900e-01  2.558e-01   2.307 0.021594 *
## I(weight^2)        1.052e-06  3.344e-07   3.146 0.001784 **
## I(displacement^2)  7.243e-05  3.324e-05   2.179 0.029954 *
## I(horsepower^2)    4.604e-04  1.331e-04   3.458 0.000605 ***
## I(year^2)          7.090e-02  1.207e-02   5.875 9.25e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.782 on 380 degrees of freedom
## Multiple R-squared:  0.8765, Adjusted R-squared:  0.873
## F-statistic: 245.3 on 11 and 380 DF,  p-value: < 2.2e-16
```

INTERPRETATION #Weight, displacement, and horsepower have a non-linear relationship with mpg #Horsepower and year are among the strongest predictors, with horsepower negatively affecting mpg #the significant F-statistic confirms that at least one predictor is meaningfully related to mpg.

```
mpg3 <- lm(log(mpg) ~ . - name, data = Auto)
summary(mpg3)
```

```
##
## Call:
## lm(formula = log(mpg) ~ . - name, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40955 -0.06533  0.00079  0.06785  0.33925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.751e+00  1.662e-01  10.533  < 2e-16 ***
## cylinders    -2.795e-02  1.157e-02  -2.415  0.01619 *
## displacement  6.362e-04  2.690e-04   2.365  0.01852 *
## horsepower   -1.475e-03  4.935e-04  -2.989  0.00298 **
## weight       -2.551e-04  2.334e-05 -10.931  < 2e-16 ***
## acceleration -1.348e-03  3.538e-03  -0.381  0.70339
## year          2.958e-02  1.824e-03  16.211  < 2e-16 ***
## origin        4.071e-02  9.955e-03   4.089 5.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1191 on 384 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8773
## F-statistic: 400.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

INTERPRETATION # the number of cylinders, horsepower, weight, and year are particularly impactful, with increasing weight #negatively affecting fuel efficiency while newer vehicles and increased displacement may contribute positively #The negative coefficient for cylinders (-0.02795) signifies tha the number of cylinders increases, fuel efficiency decreases #the negative coefficient for horsepower (-0.001475) indicates that more powerful engines tend to reduce mpg

#chapter 3 (10)

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.

```
library(ISLR)
data(Carseats)
```

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

```
##
## 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
```

INTERPRETATION #The model explains about 23.93% of the variability in sales #The coefficient for Price is -0.054459, that for every one-unit increase in price, sales are expected to decrease by about 0.054, indicating a negative relationship between price and sales. (b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

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

INTERPRETATION #The coefficient for the US variable is 1.200573, which is statistically significant (p = 4.86e-06) #The Urban variable has a coefficient of -0.021916, but it is not statistically significant (p = 0.936) (c) Write out the model in equation form, being careful to handle the qualitative variables properly. ***interpertation*** Sales^=13.043469−0.054459·Price−0.021916·Urban+1.200573·US Where:Urban = 1 for a store in an urban location, else 0 US = 1 for a store in the US, else 0

d. For which of the predictors can you reject the null hypothesis H0 : βj = 0? ***interpertation*** This question is asking about the results from the parameter T-tests. Based on the output & comments from part (b), we can reject the null hypothesis for the Price and US predictors, but there is insufficient evidence to reject the null hypothesis that the coefficient for Urban is zero.

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.

```
sales_lm_2 <- lm(Sales ~ Price + US, data = Carseats)
summary(sales_lm_2)
```

```
##
## 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?

```
summary(model)
```

```
##
## 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
```

```
summary(sales_lm_2)
```

```
##
## 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
```

```
anova(model,sales_lm_2)
```

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

interpertation For model (a), we have:R2 = 0.23928 adjusted R2 = 0.23351 For model (e), we have:R2 = 0.23926 adjusted R2 = 0.23543 In this case, both models explain ~23.9% of the variance in Sales. This is an interesting case to talk about using training R2 vs $R^{-}2$ (adjusted R2) in model selection, since model (a) has a higher R2 but lower $R^{-}2$

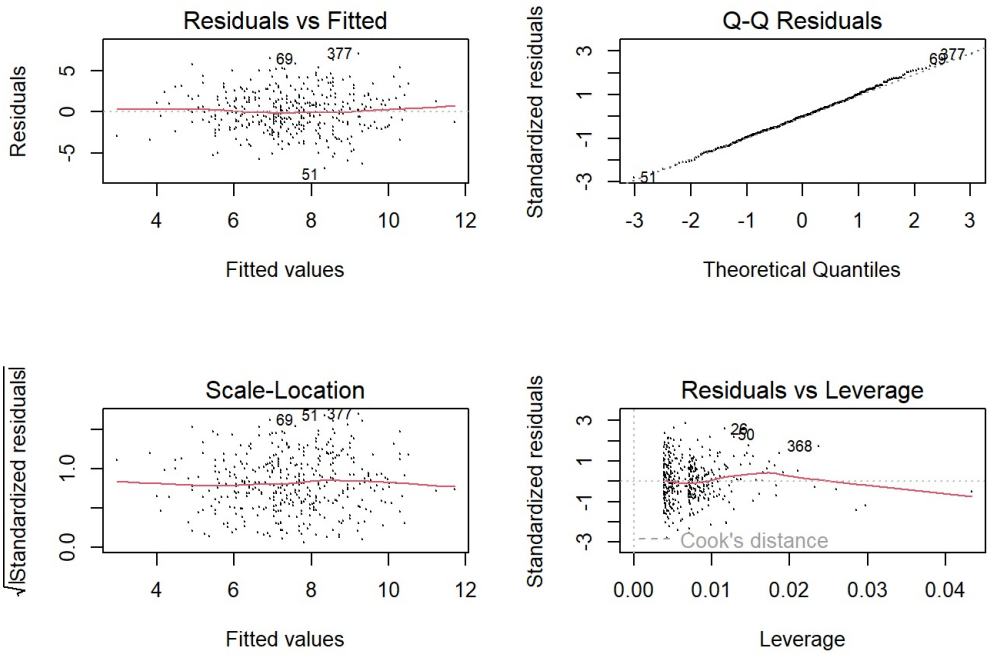
g. Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

```
confint(sales_lm_2, level = 0.95)
```

```
##                   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)?

```
par(mfrow = c(2, 2))
plot(sales_lm_2, cex = 0.2)
```



INTERPRETATION #Yes, somewhat. # a random scatter of points around the horizontal line at y=0, indicating homoscedasticity #the variance of the residuals is not constant