HW2_Q10

HW 10

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("C:/Users/Ayanna/Downloads")
cars <- read_csv("carseats.csv")
```

```
## Rows: 400 Columns: 11
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): ShelveLoc, Urban, US
## dbl (8): Sales, CompPrice, Income, Advertising, Population, Price, Age, Educ...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

10.a

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

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

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = cars)
##
## 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
```

10.b

Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative! Price = - 0.0545 Urban = - 0.0219 US = + 1.2006

Interpretation:

For every one unit increase in Price (with Urban and US factors being constant) there is a change in Sales of - 0.0545 units.

If a store is in an urban area (with Price and US factors being constant) there is a change in Sales of - 0.0219 units. That being said, since the value is greater than alpha (p = 0.05), we can conclude that there is not a statistically significant relationship between urban area and the sale of carseats (fail to reject the null).

If a store is in the US (with Price and Urban being fixed factors) there is a change in sales of 1.2006 units.

10.c

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

Sales multiple regression equation: Sales = 13.0435 - 0.0545(Price) - 0.0219(Urban) + 1.2006(US) Urban is 1 if the store IS in an urban location, otherwise 0. US is 1 if the store is in the US, otherwise 0.

10.d

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

From the information given in the previous questions, we are able to reject the null for Price and US predictors, there is enough statistically significant evidence that these factors do affect the sales of carseats. That being said, we fail to reject the null for the Urban factor as the pvalue is >0.05.

10.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.

```
new_model <- lm(Sales ~ Price + US, data = cars)
summary(new_model)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = cars)
##
## 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
```

10.f

How well do the models in (a) and (e) fit the data? model for 10.a: R^2 = 0.2393 adjusted R^2 = 0.2335

model for 10.e: R^2 = 0.2393 adjusted R^2 = 0.2354

Both models explain 23.93% of the variation occuring in the Sales of carseats in this df.

10.g

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

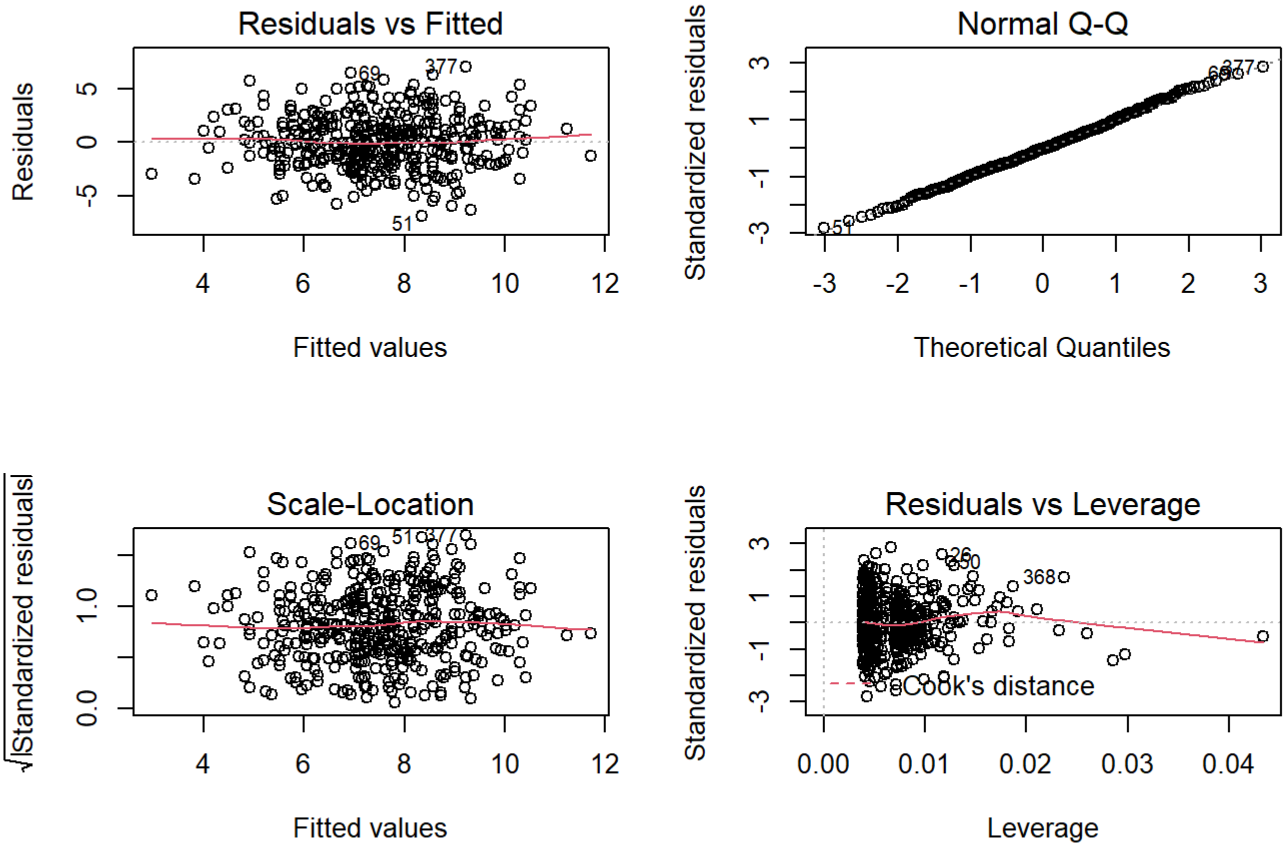
```
confint(new_model)
```

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

10.h

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

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



The residuals vs leverage plot shows that there ARE influential data points in the regression model (outliers).