Homework 1 Chapter 3:

Linear Regression: Exercise 10

==================================

10. This question should be answered using the Carseats data set.

How to read data from a URL: We can import a txt file or csv file directly into RStudio from a location on the web, by specifying the URL:

```
Carseats = read.csv("https://raw.githubusercontent.com/JWarmenhoven/ISLR-python/
                    master/Notebooks/Data/Carseats.csv", header=TRUE)
```

How to use the data directly from the ISLR R-package: You can install the ISLR package in RStudio (need to do this only once). Then, if successfully installed, you can use every time you need it by following and also access Carseats dataset directly by "attach"-ing the Carseats:

```
#install.packages("ISLR")
library(ISLR)
attach(Carseats)
```

Python: If you want to use Python, take a look at the ISLR package for Python at: https://github.com/JWarmenhoven/ISLR-python

10a.

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

```
lm.fit = lm(Sales~Price+Urban+US)
summary(lm.fit)
```

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

10b.

Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative.

Price The linear regression suggests a **significant linear** relationship between price and **average** sales, **adjusted for Urban and US status**: the p-value of the t-statistic is less than 0.05. The coefficient states a **negative linear** relationship between Price and **average** Sales: as Price increases by 1 unit, **average** Sales decreases by 54 car seats (0.0054 thousands).

Urban The linear regression suggests that **there isn't a significant linear** relationship between the location of the store and the **average** number of sales: the high p-value of the t-statistic is 0.936, and certainly greater than 0.05.

US The linear regression suggests there is a **significant linear** relationship between whether the store is in the US or not and the **average** amount of sales. The coefficient indicates a **significant positive linear** relationship between USYes and **average** Sales: if the store is in the US, the sales will increase by approximately 1201 car seats **on average**.

10c.

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

The fitted model is:

$$E(Sales_i) = 13.04 - 0.05 * Price_i - 0.02 * I(Urban = Yes)_i + 1.20 * I(US = Yes)_i$$

$$\epsilon_i \sim N(0, 2.47^2)$$

10d.

For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

Price and US variables are both significant, based on their individual p-values (both less than 0.05) – **we can reject both null hypotheses** $H_0 : \beta_{Price} = 0$ and $H_0 : \beta_{US} = 0$.

The Urban variable is not significant (its p-value is greater than 0.05), and we **fail to reject** the **null hypothesis** $H_0 : \beta_{Urban} = 0$.

10e.

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.

```
lm.fit2 = lm(Sales ~ Price + US)
summary(lm.fit2)
```

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

The fitted **reduced** model is:

$$E(Sales_i) = 13.03 - 0.05 * Price_i + 1.20 * I(US = Yes)_i$$

$$\epsilon_i \sim N(0, 2.47^2)$$

10f.

How well do the models in (a) and (e) fit the data?

Based on the F statistics and R^2 values of the two linear regressions (full and reduced), they both fit the data similarly:

The full model:

```
Multiple R-squared:  0.2393,   Adjusted R-squared:  0.2335
```

```
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

About 24% of variation ($R^2 = 0.24$) in sales is exaplained by the variation in the three predictors in the full linear regression model (Price, Urban, and US). The F-statistic is 41.52, and the p-value associated with it is less than 0.05: hence, we reject the null hypothesis that states that none of the three predictors' true effects are different from 0.

The reduced model:

```
Multiple R-squared:  0.2393,   Adjusted R-squared:  0.2354
```

```
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

About 24% of variation ($R^2 = 0.24$) in sales is still exaplained by the variation in the two predictors in the reduced linear regression model (Price and US). The F-statistic is 62.43, and the p-value associated with it is less than 0.05: hence, we reject the null hypothesis that states that none of the two predictors' true effects are different from 0.

To conclude, **both models seem to fit the data equally well**: their R^2 **values are identical**, as are **the estimated residual variances** ($\widehat{Var(\epsilon)} = 2.47^2$). Both models are explaining a significant amount of variation in the outcome variable (both F statistics' p-values are less than 0.05, and suggest that overall, both models seem to be explaining a significant amount of variation in the sales.)

10g.

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

```
confint(lm.fit2)
```

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

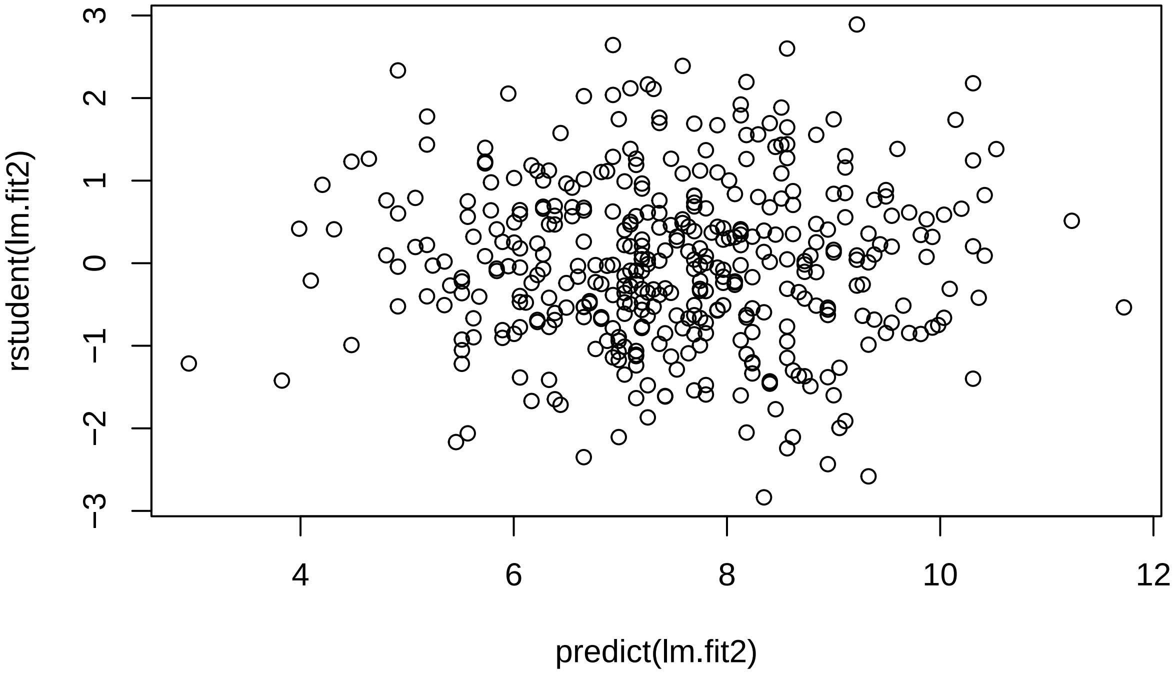
The 95% confidence interval for the true effect of Price is: **(-0.065, -0.044)**. We are 95% confident that this interval includes the true adjusted effect of Price on average Sales, adjusted for the US indicator (given the US predictor is also in the model).

The 95% confidence interval for the true effect of US is: **(0.692, 1.708)**. We are 95% confident that this interval includes the true adjusted effect of US on average Sales, adjusted for the Price variable (given the Price predictor is also in the model).

10h.

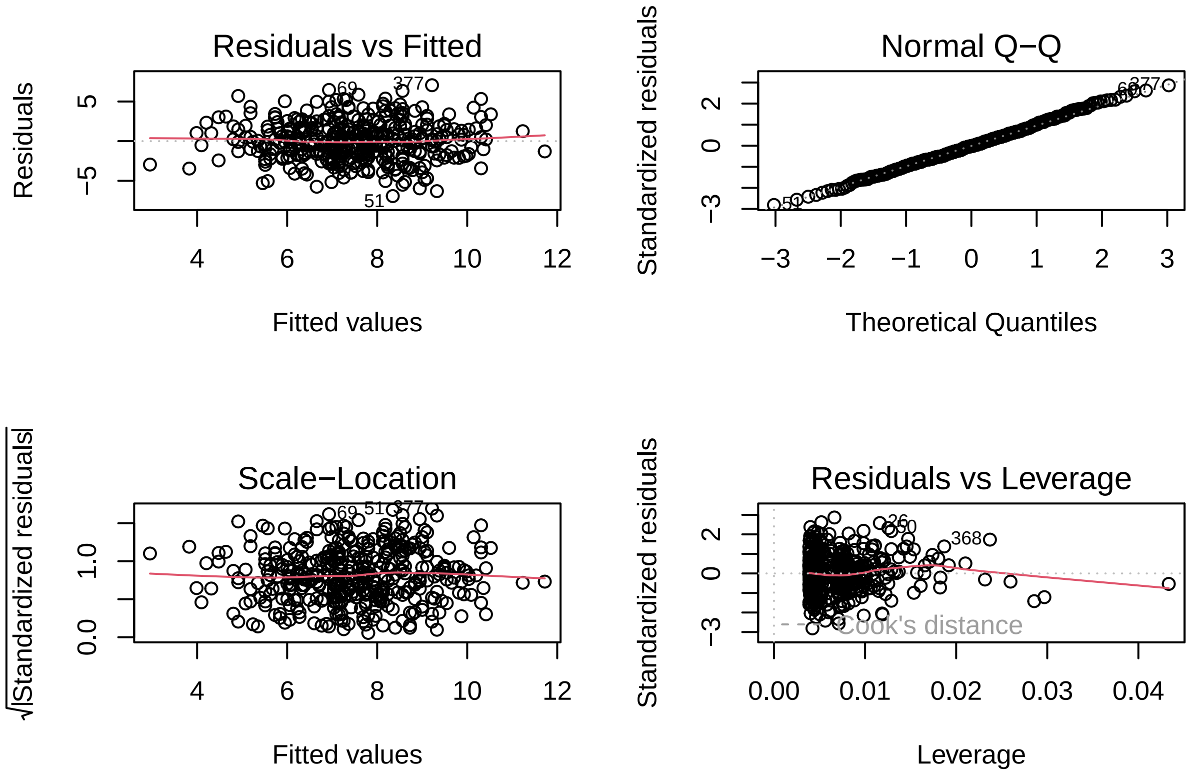
Is there evidence of outliers or high leverage observations in the model from (e)?

```
plot(predict(lm.fit2), rstudent(lm.fit2))
```



All **studentized residuals** appear to be bounded by -3 to 3, so **no potential outliers** are suggested from the linear regression. Also, **no clusters of studentized residuals** seem present (no clusters appear separated from the rest of the studentized residuals).

```
par(mfrow=c(2,2))
plot(lm.fit2)
```



There **are a few observations that exceed** $(p+1)/n$ **(3/397)** on the leverage-statistic plot that suggest that the corresponding points **have high leverage.** We do no know the impact of those high leverage points however – if we reran the regression without those high leverage points, and compared the estimates of effects (betas) before and after the deletion, we would be able to understand that impact, and see whether these high leverage points are also influential points.