Homework 4

Name: Caleb Thian Jia Le 田家樂

Student ID: NN6114035

1. Using the ”Carseats” data set to answer the following questions:

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

   b) Provide an interpretation of each coefficient in the model.

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

   d) For which of the predictors can you reject the null hypothesis $\mathrm{H}_0: \beta_{\mathrm{j}} = 0$?

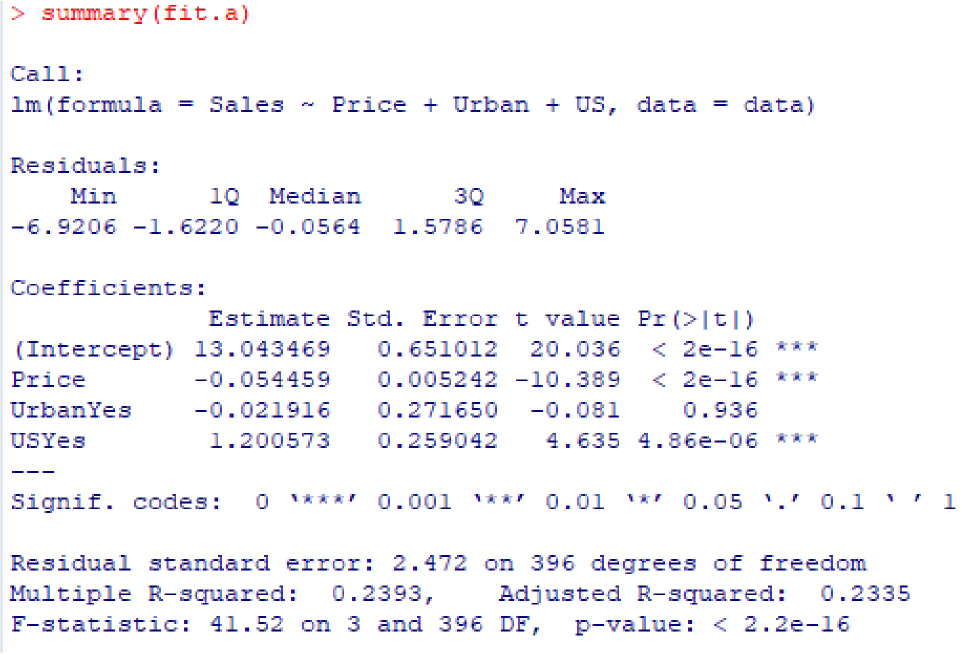
   e) Based on (d), fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

   f) How well do the models in (a) and (e) fit the data? Give the reason.

   g) Try to fit a better regression model using more predictors in data set? What is the adjusted R2 ? The analysis should provide the diagnostic figures of residuals showing the model satisfies the assumptions.

Ans:

a) Implemented in main.r.



```
> summary(fit.a)

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

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
```

$$\hat{y} = \beta_0 + \beta_1 \mathrm{Price} + \beta_2 \mathrm{Urban} + \beta_3 \mathrm{US}$$

b) The coefficients are

Features	**Coefficients**
Price	-0.05
UrbanYes	-0.02
USYes	1.20

The fitted coefficient of “Price” is -0.05, and it is significant. It means the values of Sales is affected by Price. The larger the different in Price, the more significance difference in Sales.

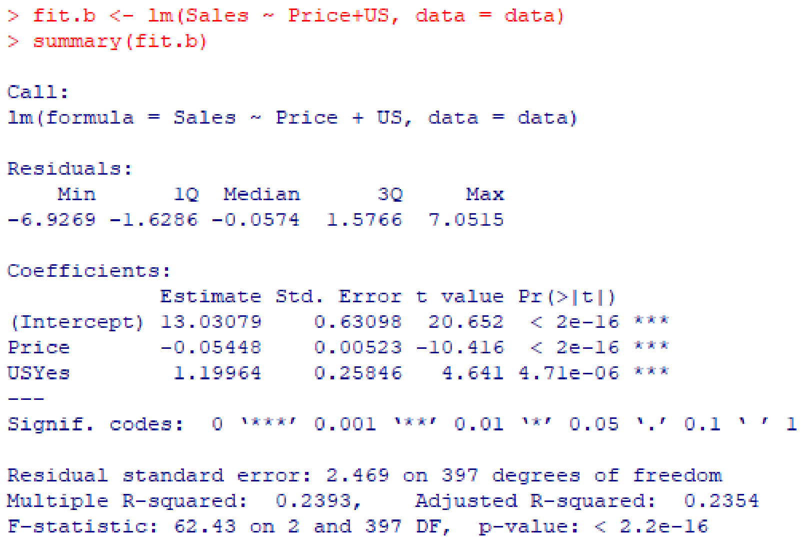
The fitted coefficient of “UrbanYes” is -0.02, but it is not significant, indicates that the values of Sales for car seats in urban and car seats in non-urban do not have significant difference. Specifically, assume that not in US, the intercept of car seats in urban is 13.04-0.02 = 13.02, and the intercept of car seats in non-urban is 13.04.

The fitted coefficient of “USYes” is 1.20, and it is significant, indicates the values of Sales for car seats in US and car seats in non-US have significant difference. Specifically, assume that not in urban, the intercept of car seats in US is 13.02+1.20 = 14.22, and the intercept of car seats in non-US is -493.73.

c)
$$\hat{y} = \begin{cases} 14.22\text{-}0.05\text{Price}, \text{when Urban} = \text{Yes and US} = \text{Yes} \\ 13.02\text{-}0.05\text{Price}, \text{when Urban} = \text{Yes and US} = \text{No} \\ 14.24\text{-}0.05\text{Price}, \text{when Urban} = \text{No and US} = \text{Yes} \\ 13.04\text{-}0.05\text{Price}, \text{when Urban} = \text{No and US} = \text{No} \end{cases}$$

d) From summary at (a), Price and US are significant variables. Thus, for these predictors, null hypothesis can be rejected.

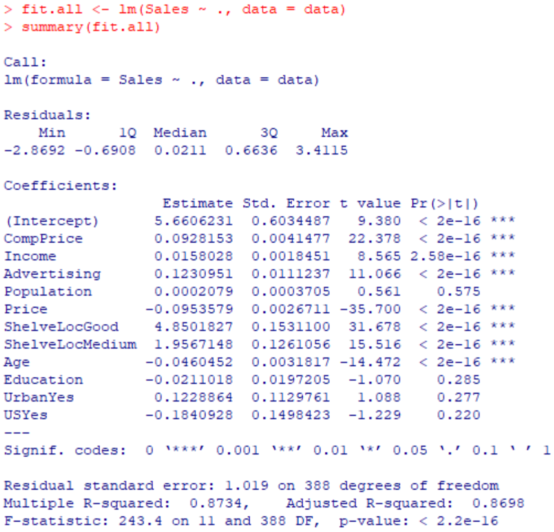
e) Fit a regression model is fitted by Price and US.



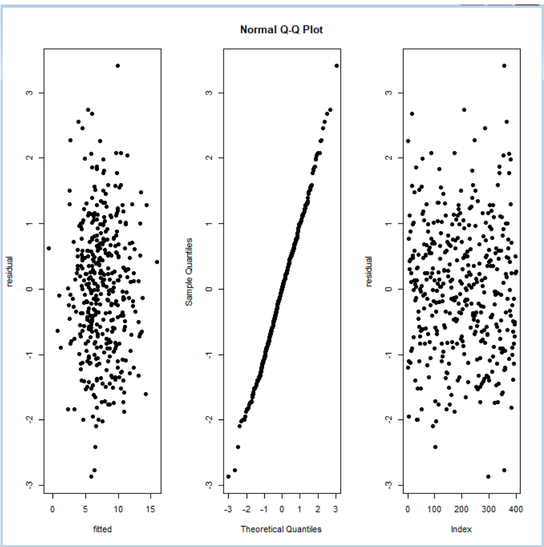
```
> fit.b <- lm(Sales ~ Price+US, data = data)
> summary(fit.b)

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

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) The model in (e) is slightly better than the model in (a) as the adjusted R^2 is 0.2354 and 0.2335 respectively, which is slightly higher. This is because the model in (e) uses less variables but only slightly reduce the value of R^2 compare with the model in (a).

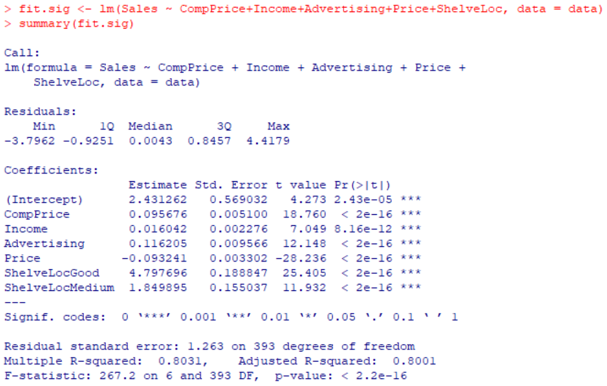
g) By fitting the regression model by all predictors, we can observe that CompPrice, Income, Advertising, Price, ShelveLoc, Age are significant variables.



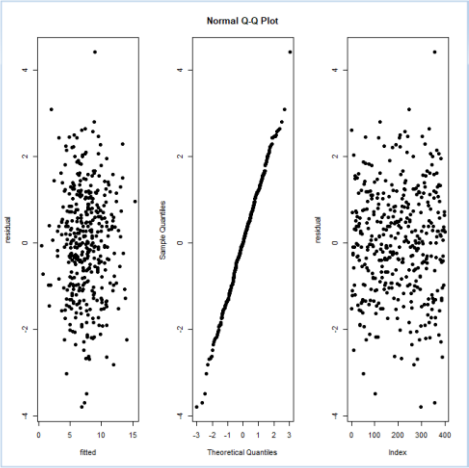
```
> fit.all <- lm(Sales ~ ., data = data)
> summary(fit.all)

Call:
lm(formula = Sales ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8692 -0.6908  0.0211  0.6636  3.4115 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
Income           0.0158028  0.0018451   8.565 2.58e-16 ***
Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
Population       0.0002079  0.0003705   0.561    0.575    
Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
Education       -0.0211018  0.0197205  -1.070    0.285    
UrbanYes         0.1228864  0.1129761   1.088    0.277    
USYes           -0.1840928  0.1498423  -1.229    0.220    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,    Adjusted R-squared:  0.8698 
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```



Fit the model with significant variables.



```
> fit.sig <- lm(Sales ~ CompPrice+Income+Advertising+Price+ShelveLoc, data = data)
> summary(fit.sig)

Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price + 
    ShelveLoc, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7962 -0.9251  0.0043  0.8457  4.4179 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      2.431262   0.569032   4.273 2.43e-05 ***
CompPrice        0.095676   0.005100  18.760  < 2e-16 ***
Income           0.016042   0.002276   7.049 8.16e-12 ***
Advertising      0.116205   0.009566  12.148  < 2e-16 ***
Price           -0.093241   0.003302 -28.236  < 2e-16 ***
ShelveLocGood    4.797696   0.188847  25.405  < 2e-16 ***
ShelveLocMedium  1.849895   0.155037  11.932  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.263 on 393 degrees of freedom
Multiple R-squared:  0.8031,    Adjusted R-squared:  0.8001 
F-statistic: 267.2 on 6 and 393 DF,  p-value: < 2.2e-16
```



Both 2 models hold the assumption of multiple regressions, which are constant variance, normality and independent. Then, the best model is using all the predictors as it has a higher adjusted R^2 = 0.8698 than the model only with significant variables which adjusted R^2=0.8001.

2. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10.$

   a) True or False

      i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

      ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

      iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough

      v. Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

   b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

   Ans:

   a) $\hat{y} = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4\text{-}10X_5 = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2\text{-}10X_1X_3$

      Assume $0 \le X_1 \le 4.3$[1]$, 55 \le X_2 \le 145$[2]

      i. Under the assumption, $y_{\text{graduate}}\text{-}y_{\text{high-school}} = 35\text{-}10X_1 \in$ [-8,35], let assume that GPA is under normal distribution, than on average, $\overline{y_{\text{graduate}}\text{-}y_{\text{high-school}}} = 35\text{-}10(2.15) = 13.5$, thus, the statement is false.

      ii. True, please refer to the explanation in i.

      iii. Under the assumption, $y_{\text{graduate}}\text{-}y_{\text{high-school}} = 35\text{-}10X_1 \in$ [-8,35], provide GPA is high enough, than $y_{\text{graduate}}\text{-}y_{\text{high-school}} < 0$, the statement is true.

[1] Refer https://zh.wikipedia.org/wiki/成績平均積點

[2] Refer https://zh.wikipedia.org/wiki/智商

iv. False, please refer to the explanation in iii.

v. $0 \le \mathrm{X}_4 = \mathrm{X}_1\mathrm{X}_2 \le 623.5$
$0 \le 0.01(\mathrm{X}_4) \le 6.235$

This indicates that the maximum of the interaction effect is 6.235, compare with the intercept term, $\frac{6.235}{50} = 0.1247$, it is very little evidence. The statement is true.

b) $\hat{y} = 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0)(110) - 10(4.0)(1) = 137.1$

The predicted salary is 137.1 thousands of dollars.