STA314 Homework 2

student number: 1003942326

Yulin WANG

11/10/2019 due

Question 1

(a)

fit the model first: let Y be the response, starting salary after graduation(in thousands of dollars)

$$\hat{Y} = 50 + 20 * GPA + 0.07 * IQ + 35 * Gender + 0.01 * GPA : IQ - 10 * GPA : Gender$$

model for male(Gender=0):

$$\hat{Y}_M = 50 + 20 * GPA + 0.07 * IQ + 0.01 * GPA : IQ$$

model for female(Gender=1):

$$\hat{Y}_F = 50 + 20 * GPA + 0.07 * IQ + 0.01 * GPA : IQ - 10 * GPA + 35$$

Since we have:

$$\hat{Y}_M - \hat{Y}_F = 10 * GPA - 35 > 0 \Rightarrow GPA > 3.5$$

So, for a fixed value of IQ and GPA, given the GPA above 3.5, males will earn more on average than females. Thus, iii is correct.

(b)

Since we have:

$$\hat{Y}_F = 50 + 20 * 4.0 + 0.07 * 110 + 0.01 * 4.0 * 110 - 10 * 4.0 + 35 = 137.1$$

So, the predicted salary of this female is 137.1 thousand dollars.

(c)

False.

- The scale of IQ is much larger than other predictors (about 100 versus 0-4 for GPA and 0-1 for Gender), so even if all predictors have the same impact on salary, coefficients will be smaller for IQ predictors.

- We need to compute the p-value for the estimate of coefficient to determine if a predictor is statistically significant or not. However, we do not have enough information(standard error of $\hat{\beta}_4$) here.

Question 2

(a)

```r
library('ISLR')
data(Carseats)
model1 <- lm(Sales ~ Price + Urban + US, data=Carseats)
summary(model1)
```

```
##
## 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
```

(b)

- Price:
  $\hat{\beta}_1 = -0.054459$; p-value$< 2e-16$; statistically significant
  interpretation: For each dollar increase in Price, Sales will decrease by about 54 on average.

- UrbanYes:
  $\hat{\beta}_2 = -0.021916$; p-value$= 0.936$; not statistically significant
  interpretation: Sales are about 22 lower on average for Urban locations.
  But since the p-value$= 0.936$, then there is no evidence to reject $H_0 : \beta_2 = 0$, so there is no relationship between Sales and whether the location is Urban or not.

- USYes:
  $\hat{\beta}_3 = 1.200573$; p-value$= 4.86e-06$; statistically significant
  interpretation: Sales are about 1,201 higher on average in the US locations.

(c)

$$\hat{Sales} = 13.043 - 0.054 * Price - 0.022 * UrbanYes + 1.201 * USYes$$

(d)

We can reject the null hypothesis for Price and USYes: $H_0 : \beta_1 = 0$ and $H_0 : \beta_3 = 0$ respectively, since their coefficients have very small p-values(much smaller than 0.05).

(e)

```
model2 <- lm(Sales ~ Price + US, data=Carseats)
summary(model2)
```

```
##
## 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)

For model1: $R^2 = 0.2393$ and $RSE = 2.472$
For model2: $R^2 = 0.2393$ and $RSE = 2.469$
Since these two models have the same R^2 value, but model2 has a smaller RSE value than model1, so we can conclude that model2 which without the variable UrbanYes fits the data better.