Homework 2

Question 1

(1) Answer III is correct. Given the conditions, we can write $\hat{Salary} = 50+20GPA+0.07IQ+35Gender+0.01GPA*IQ-10GPA*Gender$. It can be rewritten to $\hat{Salary} = 50+20GPA+(35-10GPA)Gender+0.07IQ+0.01GPA*IQ$ Notice that adjusting GPA will change the effect of Gender on Salary.

Since we can set $X_3 = 1$ if the gender is female ,and $X_3 = 0$ if the gender is male,

$\hat{Salary}(female) = 85 + 10GPA + 0.07IQ + 0.01GPA * IQ$

$\hat{Salary}(male) = 50 + 20GPA + 0.07IQ + 0.01GPA * IQ$

We can easily see that $\hat{Salary}(male) > \hat{Salary}(female)$ if GPA>3.5 Thus, holding a fixed value of IQ and GPA, males earn more on average than females, provided that the GPA is high enough (greater than 3.5).

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

(3) False. To see if an interaction has an effect on the response variable, we need to check the p-value of GPA*IQ. It is possible that p-value is smaller than some statistical threshold with the coefficient being small. If p-value is small, then there is strong evidence to reject the null hypothesis $H_0 : \beta_4 = 0$. In other words, the true relationship is not additive. We can also check R^2 for both models with and without the interaction term to see how much the variability of the response changes in both cases. On the other hand, the coefficient for the interaction term is only the estimator for the true parameter. It cannot provide evidence of the interaction effect.

Question 2

```
library(MASS)
library(ISLR)
```

```
summary(Carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75
##  Median :272.0   Median :117.0   Medium:219   Median :54.50
##  Mean   :264.8   Mean   :115.8                Mean   :53.32
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00
##  Max.   :509.0   Max.   :191.0                Max.   :80.00
##    Education    Urban       US
##  Min.   :10.0   No :118   No :142
##  1st Qu.:12.0   Yes:282   Yes:258
```

```
##  Median :14.0
##  Mean   :13.9
##  3rd Qu.:16.0
##  Max.   :18.0
```

```
attach(Carseats)
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
```

(b)

1. Ignoring all the effect, the sales of the child car seat on average is 13 units.
2. (Price) For each $1000 increase in price, there will be an average decrease in sales of 54 units.
3. (UrbanYes)The average difference of sales in between urban and non-urban is -0.0219 unit. Further we can see from the high p-value that there is no relationship between the number of sales and the location of the store.
4. (USYes)The average difference of sales between whether the store is in US or other countries is 1.20 units. If the store is in the US, the sales will increase by 1201 units.

(c)

$\hat{Sales} = 13.043 - 0.054Price - 0.022UrbanYes + 1.201USYes$ where UrbanYes=1 if the store location is in urban, else UrbanYes= 0, USYes=1 if the store location is in US, else USYes = 0

(d)

We can reject the null hypothesis for predictor Price and USYes because both of them have small p-values.

(e)

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

```r
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
```

(f)

Let's call model in (a) model 1 and model in (e) model 2. We can see from the summary that adjusted R^2 for model 1 is 0.2335, while adjusted R^2 for model 2 is 0.2354. It means model 1 explains 23.35% variance in sales, while model 2 explains 23.54% variance in sales. It is also shown that the RSE in model 1 is 2.472, and the RSE in model 2 is 2.469. Note that RSE is an absolute measure of lack of fit of the model to the data, so the better model has lower RSE. As we can see, the adjusted R-squared is higher and the RSE is lower in model 2 compared to model 1. Thus, model 2 fits the data (slightly) better.