Assignment 2

2019-10-09

Question 1

If we write down the model using the $\hat{\beta}$ coefficients we are provided, we have:

$$\hat{y} = 50 + 20 * GPA + 0.07 * IQ + 35 * Gender + 0.01 * (GPA \times IQ) - 10 * (GPA \times Gender)$$

where Gender is 1 for Female and 0 for Male and $\hat{y}$ represents salary in thousands

So, if we write separate models for male and female, we have:

For Male, the model is:

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

For Female, the model is:

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

Part a:

Here, all the options are comparing Male's Salary with respect to Female's.

So, firstly, let's just equate both the models to see how does the inequality work out. If we want to see when, on average, do female earn more than male, the inequality we have is:

$$85 + 10 * GPA + 0.07 * IQ + 0.01 * (GPA \times IQ) > 50 + 20 * GPA + 0.07 * IQ + 0.01 * (GPA \times IQ)$$
$$85 + 10 * GPA > 50 + 20 * GPA$$
$$3.5 > GPA$$

Therefore, when the GPA is smaller than 3.5, females, on average, earn more than males.

(i) is **incorrect** as males, on average, earn more than females only when GPA is greater than 3.5, and not in general for any constant value of GPA.
(ii) is **incorrect** as females, on average, earn more than males only when GPA is less than 3.5, and not in general for any constant value of GPA.
(iii) is **correct** as males, on average, earn more than females when GPA is greater than 3.5. So, if we quantify high GPA as GPA >3.5, this statement is correct.
(iv) is **incorrect** as for high GPA, females earn less than males, on average.

Part b:

Here, we want to predict the salary of a female with IQ 110 and GPA 4.0. So, here, if we plug in the values in our base model, we have:

$$\begin{aligned}\hat{Y} &= 85 + 10 * GPA + 0.07 * IQ + 0.01 * (GPA \times IQ) \\ &= 85 + 10 * 4 + 0.07 * 110 + 0.01 * (4 * 110) \\ &= 137.1\end{aligned}$$

Part c:

False. To verify if (GPA & IQ) together have an impact, we need to test $H_0 : \hat{\beta}_4 = 0$ and look at the p-value associated with a t-statistic or an F-statistic to come to a conclusion.

If we notice partial effect of IQ and its interactive term, we see that IQ's main effect is $(0.07 * IQ)$ and the effect of interactive term is $(0.01 * GPA * IQ)$. So, IQ hierarchical effect is $(GPA/7)$, which is around 57% the above case for GPA = 4.0. So, we can not ignore it before having a closer look at the respective p-value.

Question 2

Part a:

By installing the library ISLR, we already have the dataset Carseats. So, we will directly jump to fitting the model.

```
mod1 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(mod1)
```

```
##
## 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
```

Part b:

To interpret the coefficients, we first look at the categorical variables using contrast function

```
knitr::kable(contrasts(Carseats$Urban), caption = "Coding R uses in case of Urban vs Rural")
```

Table 1: Coding R uses in case of Urban vs Rural

	Yes
No	0
Yes	1

```
knitr::kable(contrasts(Carseats$US), caption = "Coding R uses in case of US vs Non US")
```

Table 2: Coding R uses in case of US vs Non US

	Yes
No	0
Yes	1

Based on the result of contrast function, we observe that person living in urban setting has value 1 for Urban variable and person living in rural setting has value 0 for the same. For US variable, R assigns value 1 if the store is in US and 0 if it is not.

Now that we know the qualitative predictors, we can start interpreting the coefficients in the model.

- Intercept represents sales(in thousands) for a store where it does not charge any price for car seats, which is in a rural setting and the store is not in US.
- For the Price coefficient, we can say that for every unit increase in the price charged for car seats, the sales decrease by approx 54.45 units(0.05445*1000) keeping all other predictors fixed.
- Coefficient for UrbanYes indicates the average difference in the sales of a store located in Urban area as compared to rural area. So, if the store is in Urban area, then keeping everything else constant, the sales go down by 21.9 units as compared to stores in Rural area. However, since the p-value is not significant, we can not be certain about this relationship.
- Lastly, the USYes coefficient can be interpreted as the average increase in Sales provided that the store is located in the United States. Thus, on average, the sales in a US store are 1200.57 units more than in a non US store keeping all other predictors remaining fixed.

Part c:

$$Sales = 13.043469 - 0.054459 * Price - 0.021916 * UrbanYes + 1.200573 * USYes + \epsilon$$

where, UrbanYes is 1 if the store is in Urban setting and 0 if not while USYes is 1 if the store is in US and 0 if not

Part d:

For rejecting the null hypothesis $H_0 : \beta_j = 0$, we would want a small p-value (less than 0.05). So, here we can reject null hypothesis for **Price** and **USYes** variable as their p-value is less than 0.05. A small p-value indicates that we observe an association between the predictor and the response. So, we reject the null hypothesis and declare a relationship to exist between predictor and response.

Part e:

On the basis of previous question, we now want to fit a new smaller model that only uses the predictors for which there is evidence of association with the response.

```r
mod2 <- lm(Sales ~ Price + US, data = Carseats)
summary(mod2)
```

```
##
## 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
```

Part f:

On comparing the output of summaries, we see that R^2 is the same for both the models. So, while going from smaller to bigger model (i.e., model in part (e) to model in part (a)), because the R^2 does not increase, we can continue with the smaller model and drop the Price$Urban variable as it is not helping improving the fit.

Moreover, while going from smaller model to base model, RSE increases, F-statistic decreases and the Adjusted R^2 decreases a bit as well. So, we can say that smaller model would be marginally better for all the above mentioned reasons and it is also easier to interpret.

But on average, if we just look at the summary statistics for model comparison, the models are quite similar.