UNIVERSITY OF MICHIGAN
DATA MINING
STATS415

ASSIGNMENT 2

Author: *Shu* ZHOU
ID: 19342932
Lab Section: 001

September 15, 2020

1 Q1.

The resulting formula is

$$y = 40 + 3 \times GPA - 2 \times Gender + 1.5 \times PriorExperience - 0.5 \times GPA \times Gender - 0.1 \times GPA \times PriorExperience \tag{1}$$

1.1

For a female employee with a GPA of 3.5 and 2 years experience, the start salary is

$$y = 40 + 3 * 3.5 - 2 + 1.5 \times 2 - 0.5 \times 3 \times 1 - 0.1 \times 3.5 \times 2 = 49.3 Kdollars \tag{2}$$

1.2

Male employees earn more on average.

Since the coefficient β_3 (for gender) and β_5 (for gender and GPA) are both negative, hence the greater the gender variable, the less the start salary y. As a result, male (with gender variable = 0) earn more than female (with gender variable = 1).

1.3

The coefficient for the interaction between GPA and gender is negative. Which means that with the the combined action of these two predictors is less then the sum of the individual effects of GPA and gender.

1.4

False, since the coefficient β_5 is not zero, it reflects some interactions between GPA and experience. Since both GPA and experience are great in great scales (GPA from 0-4 Experience can go to the decimals), the impact of this interaction might cause a great impact on the start salary.

1.5

The plot was performed by R

```
curve(40+3*x,0,4,xlab="GPA",ylab="Start Salary",bty="l",add=T,col="blue", main = "GPA vs. Start Salary")
text(1,46,expression(paste("y = ",40+3*x)),col="blue")
text(1,45,expression(paste("Male")),col="blue")
curve(38+2.5*x,0,4,bty="l",add=T,col="red", main = "female")
text(3,42,expression(paste("y = ",38+2.5*x)),col="red")
text(3,41,expression(paste("Female")),col="red")
```

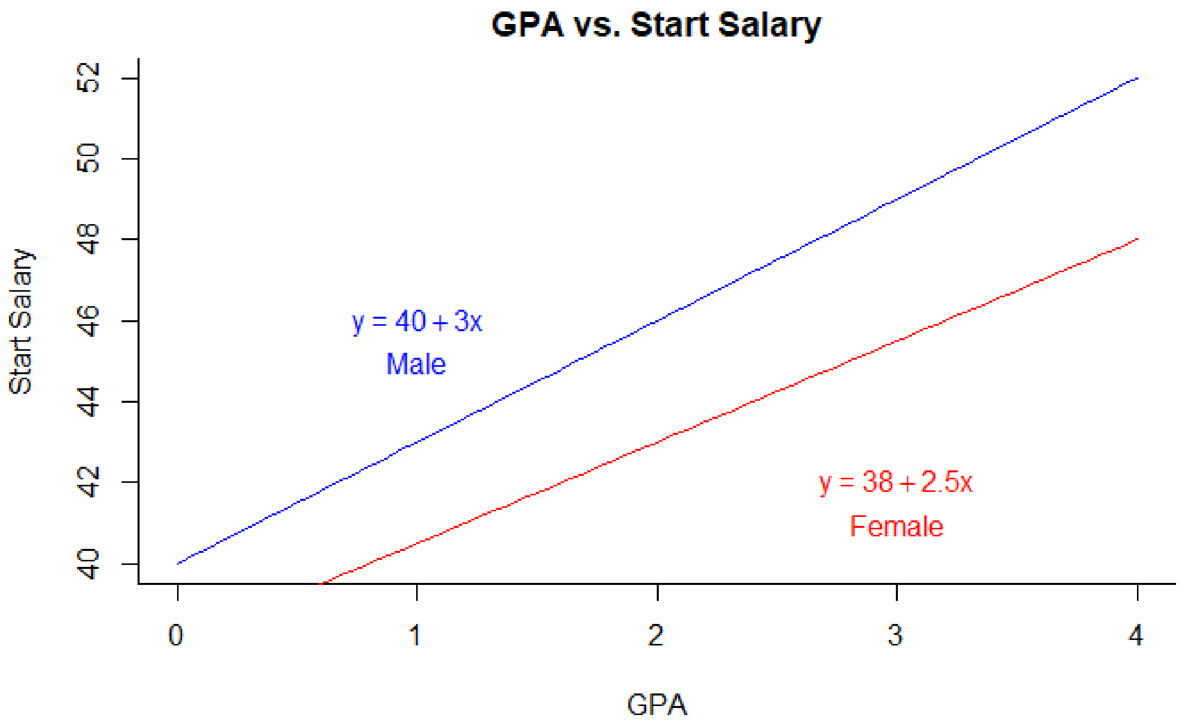


Figure 1: GPA vs. Salary, with the bl

1.6

We need to use the Hypothesis Test for Regression Slope. First, we subtract one slope from the other, and test whether the slope of the resulting curve is equal to zero. The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

Then, Use a linear regression t-test to determine whether the slope of the regression line differs significantly from zero.

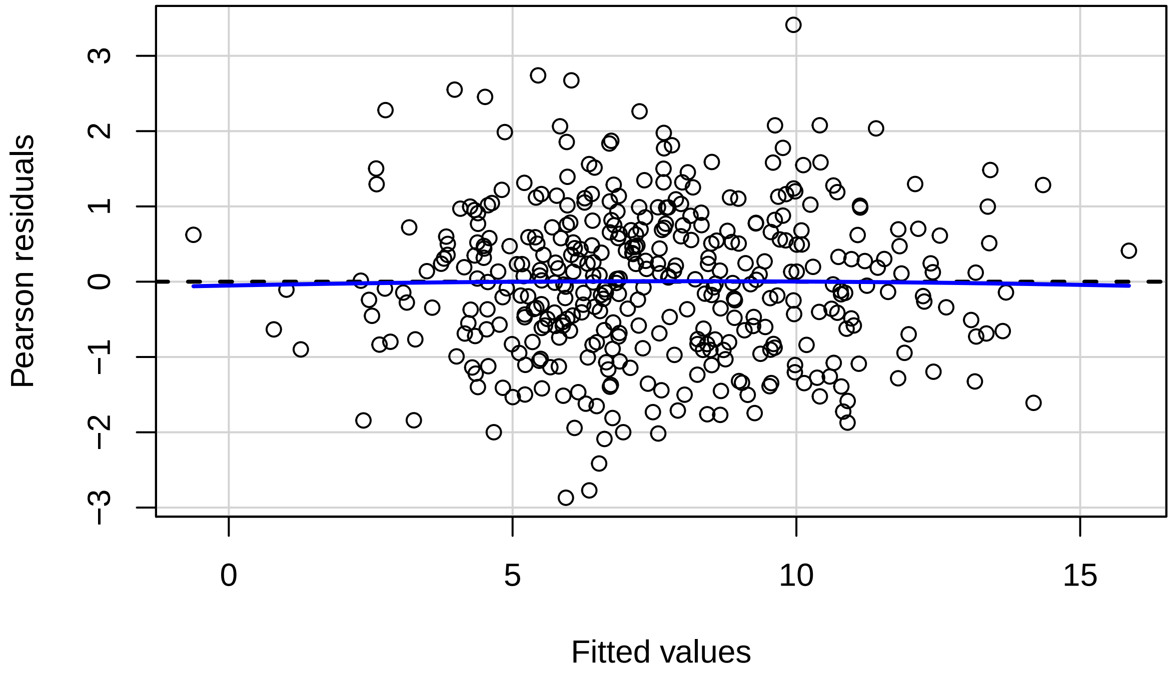
2 Q2.

(a)

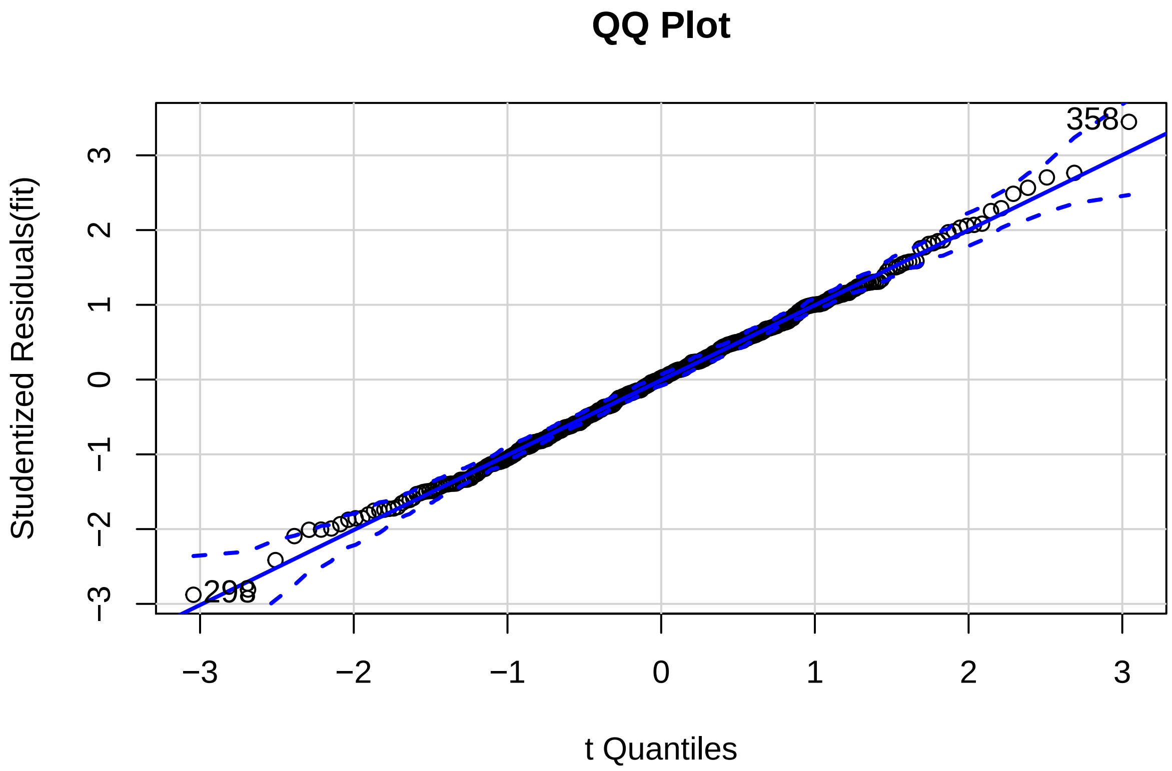
```
library(ISLR)
data("Carseats")
fit <- lm(Sales ~ CompPrice+Income+Advertising+Population+Price+ShelveLoc+Age+Education+Urban+US, data =
summary(fit)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Population +
##     Price + ShelveLoc + Age + Education + Urban + US, data = Carseats)
##
## 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
```

```
##The multiple R-squared value is 0.8698, which shows that this regression
##can interpret 87% of the changes of the dependent variable.
residualPlot(fit)          #Diagnostic residual plots
```



```
qqPlot(fit, main="QQ Plot") #qq plot for studentized residuals
```



```
## [1] 298 358
```

```
##There is no clear interaction between the predicted value and residuals
```

(b)

```
##We can see that CompPrice, Income, Advertising, Price,
##and ShelveLoc have significant p-values.

##For the variable "Urban", we have the P-value = 0.277 > 0.05, hence, we
##rejected the hypothesis that the variable Urban is significant.
```

(c)

```
fit1 <- lm(Sales ~ CompPrice+Income+Advertising+Price+ShelveLoc, data = Carseats)
summary(fit1)
```

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

```
##The multiple R-squared value is 0.8001, which shows that this regression can
##interpret 80% of the changes of the dependent variable

##The R-squared value  of the reduced model slightly decreased from the previous
##value with the full model.
```

(d)

```
anova(fit,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ CompPrice + Income + Advertising + Population + Price +
##     ShelveLoc + Age + Education + Urban + US
## Model 2: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    388 402.83
## 2    393 626.51 -5   -223.68 43.088 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The P-value is significant, i.e. we can not reject the hypothesis that the two
# models have different variance.

#Hence, the different between the R-squared value is not significant, and the
#second model is better.
```

(e)

y = 2.431 + 0.096 x CompPrice + 0.016 x Income + 0.116 x Advertising - 0.093 x Price + 4.798 (If shelveLoc = Good) + 1.850 (If ShelveLoc = Medium)

(f)

```
fit2 <- lm(Sales ~ CompPrice+Income+Advertising+Price+ShelveLoc + Price:ShelveLoc, data = Carseats)
summary(fit2)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Price:ShelveLoc, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7547 -0.9336  0.0078  0.8386  4.3561
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.964179   0.795606   2.469  0.01398 *
## CompPrice              0.095881   0.005144  18.638  < 2e-16 ***
## Income                 0.015969   0.002290   6.974 1.32e-11 ***
## Advertising            0.116309   0.009596  12.121  < 2e-16 ***
## Price                 -0.089335   0.005739 -15.567  < 2e-16 ***
## ShelveLocGood          5.353757   0.920389   5.817 1.25e-08 ***
## ShelveLocMedium        2.473173   0.774915   3.192  0.00153 **
## Price:ShelveLocGood   -0.004843   0.007752  -0.625  0.53249
## Price:ShelveLocMedium -0.005441   0.006626  -0.821  0.41205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.265 on 391 degrees of freedom
## Multiple R-squared:  0.8035, Adjusted R-squared:  0.7995
## F-statistic: 199.8 on 8 and 391 DF,  p-value: < 2.2e-16
```

```
##We can see that the interaction between Price and ShelveLoc have
##non-significant p-values, hence the interaction term is not necessary.
```

(d)

```
anova(fit1,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc
## Model 2: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Price:ShelveLoc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    393 626.51
## 2    391 625.38  2    1.1343 0.3546 0.7017
```

```
#The P-value is not significant, i.e. we canreject the hypothesis that the two
# models have different variance.

#Hence, the different between the R-squared value is significant, and the
#interaction term is not necessary.
```