Solutions to HW6

I-Chen Lee

```
library(ISLR2)
data("Carseats")
```

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.

```
fit1.a <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit1.a)
```

```
##
## 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
```

The fitted regression is

$$Sales = 13.043 - 0.05Price - 0.02Urban + 1.20US,$$

where $Urban = 1$ if the store is in an urban and $Urban = 0$ if the store is not in an urban, and $US = 1$ if the store is in the US and $US = 0$ if the store is not in the US.

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

- Intercept: if the value of Price is zero and the store is not in an urban and in the US, then the expected values of sales is 13.043.
- Coefficient of Price: Given the same settings of store locations, a one-unit increase in the price is associated with a decrease in the sales by 0.05 units.
- Coefficient of Urban: Given the same settings of price and store location in US or not, the average difference on the sales between the store in the urban and that in the rural locations is -0.02 units. The sales in the rural locations is more than that in the urban locations.
- Coefficient of US: Given the same settings of price and store location in the urban or not, the average difference on the sales between the store in the US and that outside the US is 1.20 units. The sales in the US is more than that outside the US.

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

$$\text{Sales} = \begin{cases} 13.043 - 0.05\ \text{Price} & \text{if (Urban, US)} = \text{(No, No)}, \\ (13.043 - 0.02) - 0.05\ \text{Price} & \text{if (Urban, US)} = \text{(Yes, No)}, \\ (13.043 + 1.20) - 0.05\ \text{Price} & \text{if (Urban, US)} = \text{(No, Yes)}, \\ (13.043 - 0.02 + 1.20) - 0.05\ \text{Price} & \text{if (Urban, US)} = \text{(Yes, Yes)}. \end{cases}$$

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

According to the results in (a), the p-values of Price and US are smaller than 0.05. Then, reject the null hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_3 = 0$.

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

The fitted regression with the significant variables is

$$Sales = 13.043 - 0.05 Price + 1.20 US,$$

where $US = 1$ if the store is in the US and $US = 0$ if the store is not in the US.

```
fit1.e <- lm(Sales ~ Price + US, data = Carseats)
summary(fit1.e)
```

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

The values of the adjusted R^2 of both models are smaller, and the models in (a) and (e) do not fit the data well.

(g) Try to fit a better regression model using more predictors in data set? What is the adjusted R^2? The analysis should provide the diagnostic figures of residuals showing the model satisfies the assumptions.

From the following results, a better model is

$$\text{Sales} = 5.48 + 0.09\ \text{CompPrice} + 0.02\ \text{Income} + 0.12\ \text{Advertising} - 0.10\ \text{Price} +$$
$$4.84\ \text{ShelveLocGood} + 1.95\ \text{ShelveLocMedium} - 0.05\ \text{Age},$$

where

$$\text{ShelveLocGood} = \begin{cases} 1, \text{if the quality of the shelving location is Good,} \\ 0, \text{if the quality of the shelving location is not Good,} \end{cases}$$

and

$$\text{ShelveLocMedium} = \begin{cases} 1, \text{if the quality of the shelving location is Medium,} \\ 0, \text{if the quality of the shelving location is not Medium.} \end{cases}$$

The adjusted R^2 is 0.8697 and the residual plots do not show unusual patterns. Hence, the model fits better.

```
#pairs(Carseats[,c(1:6, 8, 9)], pch = 19)
fit1.g <- lm(Sales~., data = Carseats)
summary(fit1.g)
```

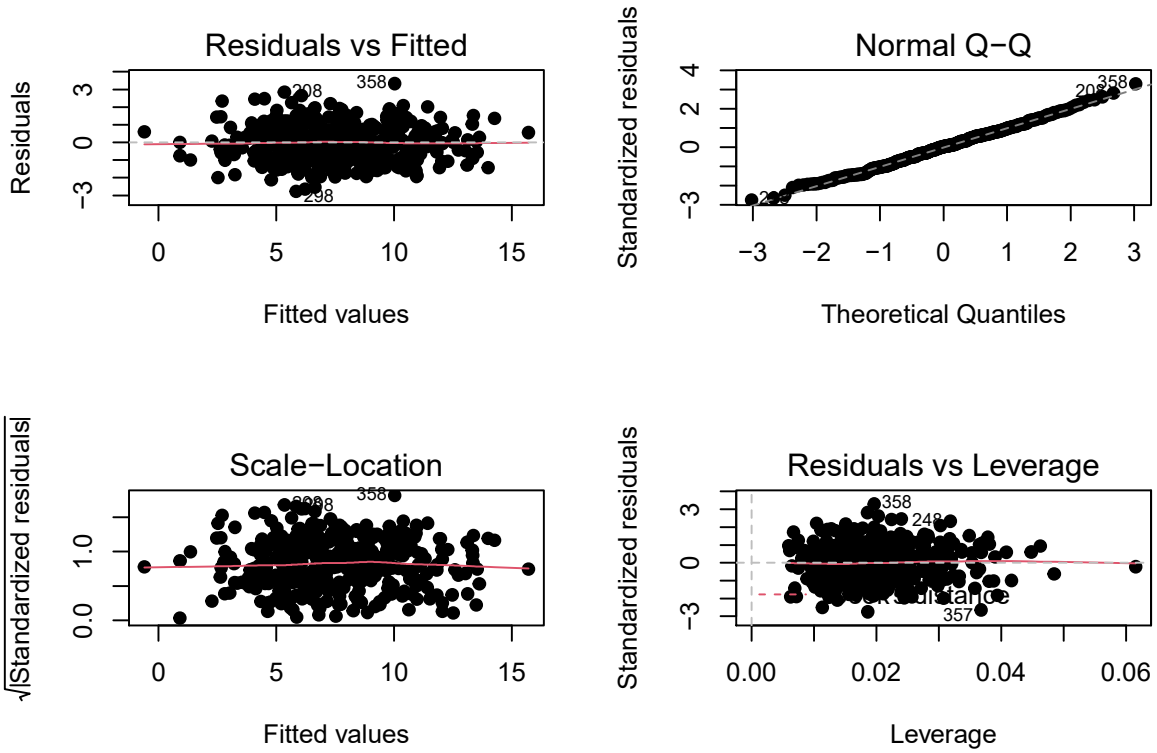
```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## 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
```

```
fit1.g1 <- lm(Sales ~ CompPrice + Income + Advertising + Price +
                    ShelveLoc + Age, data = Carseats)
summary(fit1.g1)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7728 -0.6954  0.0282  0.6732  3.3292
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.475226   0.505005   10.84   <2e-16 ***
## CompPrice        0.092571   0.004123   22.45   <2e-16 ***
## Income           0.015785   0.001838    8.59   <2e-16 ***
## Advertising      0.115903   0.007724   15.01   <2e-16 ***
## Price           -0.095319   0.002670  -35.70   <2e-16 ***
## ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
## ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
## Age             -0.046128   0.003177  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 392 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8697
## F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(fit1.g1, pch = 19)
```



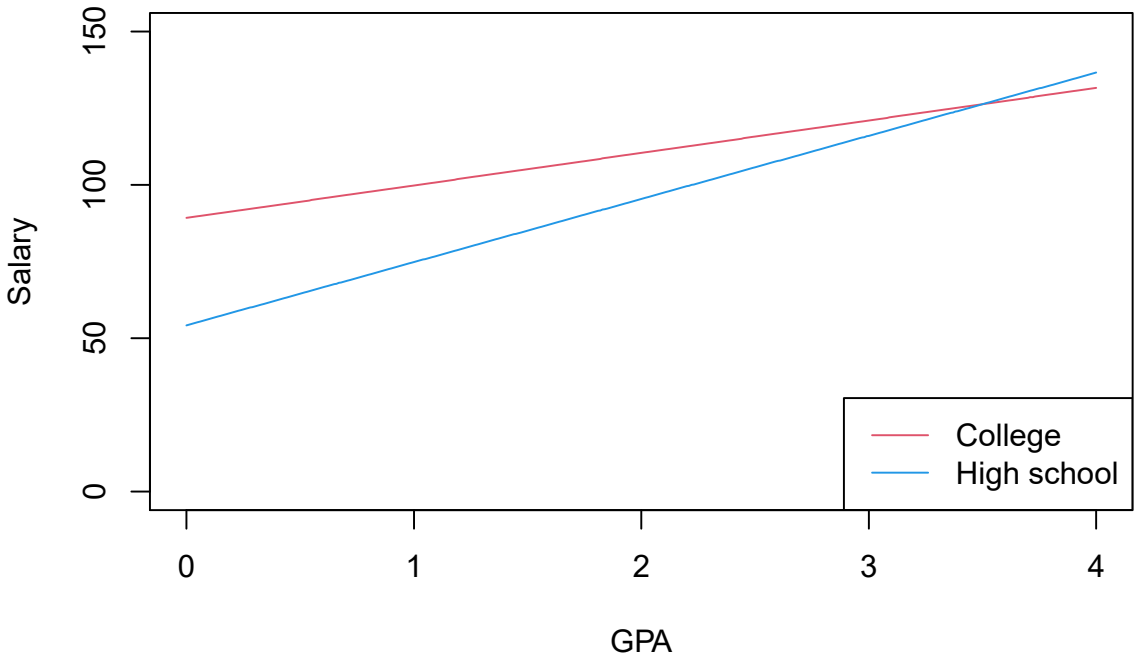
2. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, and $\hat{\beta}_5 = -10$.

$$\text{Salary} = (50 + 35) + (20 - 10)\ \text{GPA} + 0.07\ \text{IQ} + 0.01\ \text{GPA} \times\ \text{IQ}, \quad \text{for college,}$$

and

$$\text{Salary} = 50 + 20\ \text{GPA} + 0.07\ \text{IQ} + 0.01\ \text{GPA} \times\ \text{IQ}, \quad \text{for high school.}$$

```
IQ <- 60
GPA <- seq(0.0, 4, 0.01)
line.college <- 85 + 10*GPA + 0.07*IQ + 0.01* GPA * IQ
line.high <- 50 + 20*GPA + 0.07*IQ + 0.01* GPA * IQ
plot(GPA, line.college, type = "l", col = 2, ylim = c(0, 150), ylab = "Salary")
lines(GPA, line.high, col = 4)
legend("bottomright", c("College", "High school"), col = c(2, 4), lty = c(1, 1))
```



(a) True of False

Given GPA and IQ, the models are

$$\text{Salary} = (50 + 35 + 0.07\ \text{IQ}) + (20 - 10 + 0.01 \times \text{IQ})\text{GPA}, \quad \text{for college,}$$

and

$$\text{Salary} = (50 + 0.07\ \text{IQ}) + (20 + 0.01 \times \text{IQ})\ \text{GPA}, \quad \text{for high school.}$$

The difference on the intercept between two groups is 35, and the difference on the slope between two groups is -10. The cross point of two lines is at GPA $= 3.5$. If GPA ≥ 3.5, then high school graduates earn more, on average, than college graduates. If GPA ≤ 3.5, then college graduates graduates earn more, on average, than high school graduates.

i. FALSE, need to consider the interaction between GPA and levels.
ii. FALSE, need to consider the interaction between GPA and levels.
iii. TRUE, shown by the mathematical equations and the figure.
iv. FALSE, (The correct description is at iii.)
v. FALSE, need to use the hypothesis testing if it is really smaller enough.

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

$$\text{Salary} = (50+35)+(20-10)\times 4+0.07\times 110+0.01\times 4\times 110 = 137.1(\text{in thousands of dollars}), \quad \text{for college.}$$