Homework 2

Rento Saijo

Department of Mathematics, Connecticut College

STA336: Statistical Machine Learning

Yan Zhuang, Ph.D.

February 6, 2026

Disclosure

ChatGPT-5.2 was used to create the `YAML` portion and some `LaTeX` code to format the text/equations nicely; I looked into the documentation for each of the package it used and added/removed unnecessary formattings. See the original `RMD` file here.

Problem 1 (a)

Given estimates

$$\hat{\beta}_0 = 50, \quad \hat{\beta}_1 = 20, \quad \hat{\beta}_2 = 0.07, \quad \hat{\beta}_3 = 35, \quad \hat{\beta}_4 = 0.01, \quad \hat{\beta}_5 = -10,$$

the fitted regression function is

$$\hat{Y} = 50 + 20\,\text{GPA} + 0.07\,\text{IQ} + 35\,\text{Level} + 0.01(\text{GPA} \cdot \text{IQ}) - 10(\text{GPA} \cdot \text{Level}).$$

For fixed GPA and IQ, let us compare predicted salary for college versus high school:

$$\hat{Y}_C = 50 + 20\,\text{GPA} + 0.07\,\text{IQ} + 35 + 0.01(\text{GPA} \cdot \text{IQ}) - 10\,\text{GPA},$$

$$\hat{Y}_{HS} = 50 + 20\,\text{GPA} + 0.07\,\text{IQ} + 0.01(\text{GPA} \cdot \text{IQ}).$$

The difference between them is

$$\hat{Y}_C - \hat{Y}_{HS} = 35 - 10\,\text{GPA}.$$

College earns more when

$$\hat{Y}_C > \hat{Y}_{HS} \quad \Longleftrightarrow \quad \hat{Y}_C - \hat{Y}_{HS} > 0 \quad \Longleftrightarrow \quad 35 - 10\,\text{GPA} > 0.$$

Solve:

$$35 > 10\,\text{GPA} \quad \Longleftrightarrow \quad \frac{35}{10} > \text{GPA} \quad \Longleftrightarrow \quad 3.5 > \text{GPA}.$$

Thus:

$$\text{If GPA} < 3.5,\ \hat{Y}_C > \hat{Y}_{HS}; \qquad \text{if GPA} > 3.5,\ \hat{Y}_{HS} > \hat{Y}_C.$$

Performing the basic algebra shown above, we see that $\hat{Y}_C > \hat{Y}_{HS}$ when GPA < 3.5, and $\hat{Y}_{HS} > \hat{Y}_C$ when

GPA > 3.5. Therefore, (iii) is the correct statement: high school graduates earn more than college graduates provided that GPA is high enough (specifically, GPA > 3.5).

Problem 1 (b)

For a college graduate with IQ $= 110$ and GPA $= 4.0$:

$$\begin{aligned}\hat{Y} &= 50 + 20(4) + 0.07(110) + 35(1) + 0.01(4 \cdot 110) - 10(4 \cdot 1) \\ &= 50 + 80 + 7.7 + 35 + 4.4 - 40 \\ &= 137.1.\end{aligned}$$

Therefore, the predicted starting salary is

$$\boxed{\hat{Y} = 137.1 \text{ (thousand dollars)} = \$137{,}100.}$$

Problem 1 (c)

False. The numerical size of an interaction coefficient cannot, by itself, be used to judge whether an interaction is present. First, the practical impact of the interaction depends on the scale of the predictors: since the interaction term is GPA $\cdot$ IQ and IQ values are often around 100, the term (contribution)

$$\hat{\beta}_4(\text{GPA} \cdot \text{IQ}) = 0.01(\text{GPA} \cdot \text{IQ})$$

can be nontrivial. For example, at GPA $= 4$ and IQ $= 110$,

$$0.01(4 \cdot 110) = 4.4,$$

which corresponds to \$4,400 in predicted salary (because Y is measured in thousands of dollars). (I could certainly use an extra few thousand dollars lol.) Second, the "statistical evidence" for an interaction effect (or any effect) is assessed using inference for β_4, such as a hypothesis test

$$H_0 : \beta_4 = 0 \quad \text{vs.} \quad H_A : \beta_4 \neq 0,$$

which yields a t-statistic and corresponding p-value, or equivalently by checking whether a confidence interval for β_4 includes 0. All in all, even a coefficient that appears "small" could be statistically significant and

practically meaningful, while a larger coefficient could fail to be significant if its standard error is large.

Problem 2 (a)

In the set-up chunk, tikZ was set to produce LaTeX figures, all necessary libraries were loaded, and the seed (123) was set. Let us load the Carseats data and check the shape and data types.

```
# Load data.
data(Carseats)


# Check data types and # of observations.
tibble::glimpse(Carseats)
```

```
## Rows: 400
## Columns: 11
## $ Sales       <dbl> 9.50, 11.22, 10.06, 7.40, 4.15, 10.81, 6.63, 11.85, 6.54, ~
## $ CompPrice   <dbl> 138, 111, 113, 117, 141, 124, 115, 136, 132, 132, 121, 117~
## $ Income      <dbl> 73, 48, 35, 100, 64, 113, 105, 81, 110, 113, 78, 94, 35, 2~
## $ Advertising <dbl> 11, 16, 10, 4, 3, 13, 0, 15, 0, 0, 9, 4, 2, 11, 11, 5, 0, ~
## $ Population  <dbl> 276, 260, 269, 466, 340, 501, 45, 425, 108, 131, 150, 503,~
## $ Price       <dbl> 120, 83, 80, 97, 128, 72, 108, 120, 124, 124, 100, 94, 136~
## $ ShelveLoc   <fct> Bad, Good, Medium, Medium, Bad, Bad, Medium, Good, Medium,~
## $ Age         <dbl> 42, 65, 59, 55, 38, 78, 71, 67, 76, 76, 26, 50, 62, 53, 52~
## $ Education   <dbl> 17, 10, 12, 14, 13, 16, 15, 10, 10, 17, 10, 13, 18, 18, 18~
## $ Urban       <fct> Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Ye~
## $ US          <fct> Yes, Yes, Yes, Yes, No, Yes, No, Yes, No, Yes, Yes, Yes, N~
```

The dataset has 400 observations and 11 variables. Most predictors are numeric while `ShelveLoc` is a categorical factor with levels like “Bad”, “Medium”, and “Good” and `Urban` and `US` appear as binary categorical factors. Let us check for any missingness.

```
# Check NAs.
colSums(is.na(Carseats))
```

```
##       Sales   CompPrice      Income Advertising  Population       Price
##           0           0           0           0           0           0
##   ShelveLoc         Age   Education       Urban          US
```

```
##           0           0           0           0           0
```

The dataset has no missing values in any column. Let us split the data into training and testing sets.

```
split <- caret::createDataPartition(Carseats$Sales, p = 0.5, list = FALSE)
train <- Carseats[split, ]
test  <- Carseats[-split, ]
```

Let us a fit a linear regression to predict `Sales` using `Price`, `Urban`, and `US` on the training set.

```
# Fit full model.
m1 <- lm(Sales ~ Price + Urban + US, data = train)
```

Prblem 2 (b)

Let us check the coefficients.

```
# Check full model.
summary(m1)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6713 -1.5836 -0.0612  1.3099  7.1117
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.426944   0.903928  14.854  < 2e-16 ***
## Price       -0.054645   0.007302  -7.484 2.35e-12 ***
## UrbanYes    -0.381431   0.390677  -0.976  0.33010
## USYes        1.140111   0.380002   3.000  0.00305 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.491 on 197 degrees of freedom
## Multiple R-squared:  0.2405, Adjusted R-squared:  0.2289
## F-statistic: 20.79 on 3 and 197 DF,  p-value: 9.539e-12
```

Problem 2 (b)

From the fitted model

$$\widehat{\text{Sales}} = \hat{\beta}_0 + \hat{\beta}_1 \,\text{Price} + \hat{\beta}_2 \,\mathbf{1}\{\text{Urban} = \text{Yes}\} + \hat{\beta}_3 \,\mathbf{1}\{\text{US} = \text{Yes}\},$$

the estimated coefficients are

$$\hat{\beta}_0 = 13.426944, \quad \hat{\beta}_1 = -0.054645, \quad \hat{\beta}_2 = -0.381431, \quad \hat{\beta}_3 = 1.140111,$$

where the baseline levels are Urban = No and US = No.

- **Intercept ($\hat{\beta}_0 = 13.426944$):** When Price = 0 and the store is in a non-urban area and outside the US, the predicted sales are 13.426944 thousand units. Note that Price = 0 is not realistic, so the intercept is merely a baseline.
- **Price ($\hat{\beta}_1 = -0.054645$):** Holding Urban and US constant, a unit (I'm assuming a dollar) increase in Price is associated with a decrease of 0.054645 thousand predicted Sales.
- **UrbanYes ($\hat{\beta}_2 = -0.381431$):** Holding Price and US constant, stores in an urban location are predicted to have 0.381431 thousand less sales than stores in a non-urban location.
- **USYes ($\hat{\beta}_3 = 1.140111$):** Holding Price and Urban constant, stores in the US are predicted to have 1.140111 thousand more sales than stores outside the US.

Problem 2 (c)

Using dummy variables for the qualitative predictors,

$$D_{\text{Urban}} = \mathbf{1}\{\text{Urban} = \text{Yes}\}, \qquad D_{\text{US}} = \mathbf{1}\{\text{US} = \text{Yes}\},$$

(with baseline levels Urban = No and US = No), the fitted model is

$$\widehat{\text{Sales}} = 13.426944 - 0.054645 \,\text{Price} - 0.381431 \, D_{\text{Urban}} + 1.140111 \, D_{\text{US}}.$$

Equivalently, written in piece-wise form:

$$\widehat{\text{Sales}} = \begin{cases} 13.426944 - 0.054645\,\text{Price}, & \text{Urban} = \text{No},\ \text{US} = \text{No}, \\ 13.426944 - 0.054645\,\text{Price} - 0.381431, & \text{Urban} = \text{Yes},\ \text{US} = \text{No}, \\ 13.426944 - 0.054645\,\text{Price} + 1.140111, & \text{Urban} = \text{No},\ \text{US} = \text{Yes}, \\ 13.426944 - 0.054645\,\text{Price} - 0.381431 + 1.140111, & \text{Urban} = \text{Yes},\ \text{US} = \text{Yes}. \end{cases}$$

Problem 2 (d)

To test each predictor, we consider (for each j) the hypotheses

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_A : \beta_j \neq 0.$$

Using the individual t-tests reported in Problem 2 (b) and using $\alpha = 0.05$:

- **Price:** $p = 2.35 \times 10^{-12} \leq \alpha$. We reject H_0 (i.e., Price is a significant predictor of Sales).
- **UrbanYes:** $p = 0.331 > \alpha$. We fail to reject H_0 (i.e., Urban is not a significant predictor of Sales).
- **USYes:** $p = 0.00305 \leq \alpha$. We reject H_0 (i.e., US is a significant predictor of Sales).

Problem 2 (e)

From part (d), the predictors with evidence of association with Sales (at $\alpha = 0.05$) are Price and US (while Urban is not). Therefore, we fit the reduced model

$$\text{Sales} \sim \text{Price} + \text{US}$$

using the training set.

```
# Fit reduced model.
m2 <- lm(Sales ~ Price + US, data = train)
summary(m2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = train)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5282 -1.5329 -0.0206  1.3217  7.0138
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.216705   0.877801  15.057  < 2e-16 ***
## Price       -0.054831   0.007298  -7.513 1.95e-12 ***
## USYes        1.083926   0.375575   2.886  0.00433 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.491 on 198 degrees of freedom
## Multiple R-squared:  0.2368, Adjusted R-squared:  0.2291
## F-statistic: 30.72 on 2 and 198 DF,  p-value: 2.394e-12
```

Problem 2 (f)

To compare the full model from (a),

$$m_1: \text{ Sales} \sim \text{Price} + \text{Urban} + \text{US},$$

to the reduced model from (e),

$$m_2: \text{ Sales} \sim \text{Price} + \text{US},$$

we use a partial F-test of

$$H_0 : \beta_{\text{UrbanYes}} = 0 \quad \text{vs.} \quad H_A : \beta_{\text{UrbanYes}} \neq 0.$$

```
# Perform partial F-test: reduced vs. full.
anova(m2, m1)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + US
```

```
## Model 2: Sales ~ Price + Urban + US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    198 1228.6
## 2    197 1222.7  1    5.9161 0.9532 0.3301
```

Since $p = 0.3301 > \alpha = 0.05$, we fail to reject H_0. Therefore, there is no evidence that including `Urban` significantly improves the model beyond `Price` and `US`; the reduced model is sufficient.

Problem 2 (g)

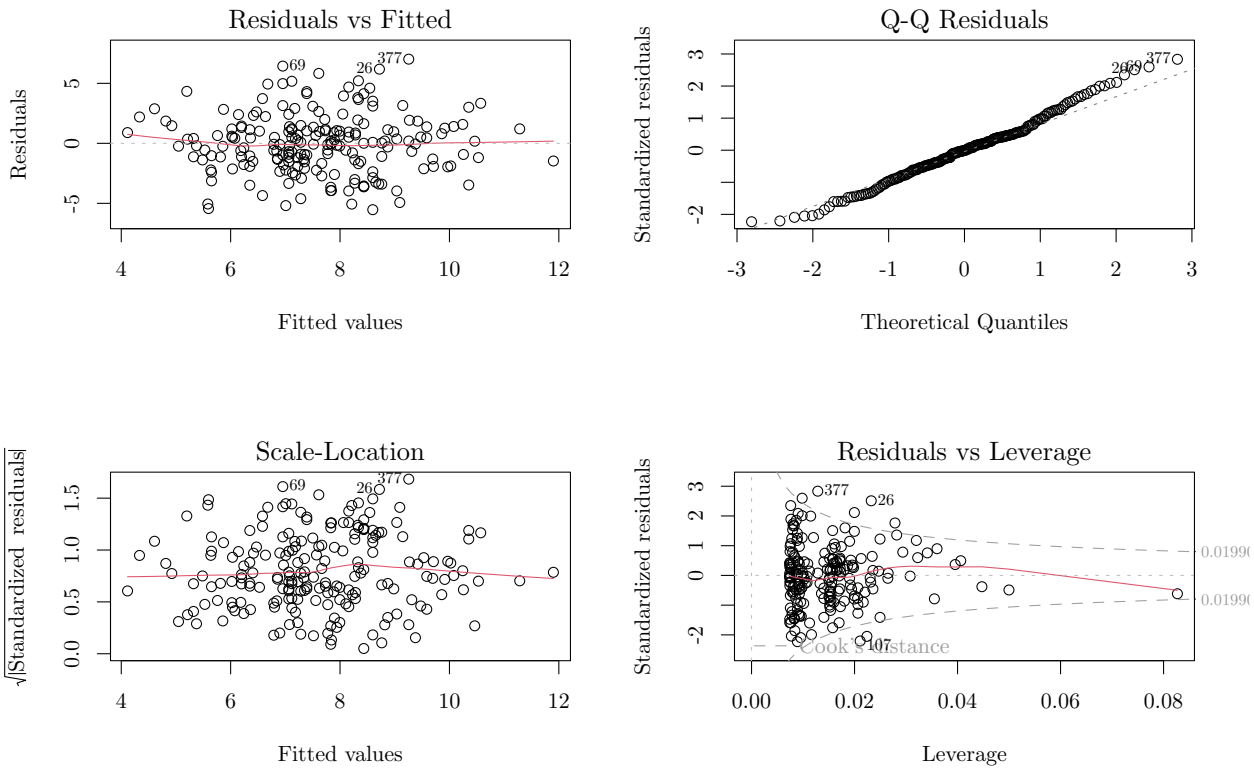
```
# Check 95% confidence interval for reduced model coefficients.
confint(m2, level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.48566685 14.94774365
## Price       -0.06922353 -0.04043835
## USYes        0.34328588  1.82456531
```

- **Intercept:** [11.486, 14.947]. When Price = 0 and US = No, the mean sales are estimated to lie between 11.486 and 14.947 thousand units. (As before, Price = 0 is not practically meaningful, so this is merely a baseline.)
- **Price:** [−0.0692, −0.0404]. Holding US fixed, increasing Price by 1 unit (dollar) is associated with a decrease in mean sales between 0.0404 and 0.0692 thousand units. Since the interval is entirely negative, Price is statistically significant at the 5% level.
- **USYes:** [0.343, 1.825]. Holding Price fixed, stores in the US (US = Yes) have mean sales that are higher by between 0.343 and 1.825 thousand units compared to stores outside the US. Since the interval is entirely positive, US is statistically significant at the 5% level.

Problem 2 (h)

```
# Check assumption plots.
par(mfrow = c(2, 2))
plot(m2, cook.levels = 4/nobs(m2))
```




```
par(mfrow = c(1, 1))
```

Using the four standard diagnostic plots for the reduced model:

- **Linearity:** the residuals are roughly centered around 0 across the range of fitted values, and the loess smoother is close to flat. There is no strong systematic pattern, though there are a few observations with relatively large positive residuals (e.g., the labeled points such as 69, 26, 377).
- **Normality:** the Q-Q plot is quite linear in the middle, but shows noticeable departures in the tails, especially in the upper tail (the labeled points 26, 69, and 377 lie above the reference line). This suggests the error distribution is approximately normal in the center but has heavier-than-normal tails / potential outliers.
- **Constant variance:** the Scale-Location plot does not show a clear funnel shape. The red curve rises slightly and then declines, indicating at most mild heteroskedasticity, but the variance appears fairly stable overall.
- **Outliers, leverage, and influence:** with Cook's distance contours displayed at $D \approx 0.0199$, most points lie well within the contour. There is one high-leverage observation far to the right (leverage around 0.08), but its standardized residual is small (around -0.5), so it does not appear highly influential. The labeled points 26 and 377 have larger standardized residuals with moderate leverage, making

them worth a look, but they are not severely beyond the Cook's distance contour.

Overall, the model assumptions look broadly reasonable: linearity and constant variance are not badly violated, normality is mostly acceptable but with tail deviations, and there is no clear evidence of extremely influential observations (though a few points warrant inspection).

Problem 2 (i)

Let us extend the reduced model by including an interaction between `Price` and the US indicator. Let

$$D_{\text{US}} = \mathbf{1}\{\text{US} = \text{Yes}\},$$

so $D_{\text{US}} = 0$ if US = No and $D_{\text{US}} = 1$ if US = Yes. The fitted interaction model has the form

$$\widehat{\text{Sales}} = \beta_0 + \beta_1 \,\text{Price} + \beta_2 \, D_{\text{US}} + \beta_3(\text{Price} \cdot D_{\text{US}}).$$

```
# Fit interaction model.
m3 <- lm(Sales ~ Price * US, data = train)
summary(m3)
```

```
##
## Call:
## lm(formula = Sales ~ Price * US, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5297 -1.5313 -0.0205  1.3192  7.0110
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.1985009  1.3309320   9.917  < 2e-16 ***
## Price       -0.0546694  0.0114905  -4.758 3.78e-06 ***
## USYes        1.1150422  1.7477159   0.638    0.524
## Price:USYes -0.0002717  0.0149025  -0.018    0.985
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.497 on 197 degrees of freedom
## Multiple R-squared:  0.2368, Adjusted R-squared:  0.2252
## F-statistic: 20.38 on 3 and 197 DF,  p-value: 1.523e-11
```

The estimated coefficients are

$$\hat{\beta}_0 = 13.1985009, \quad \hat{\beta}_1 = -0.0546694, \quad \hat{\beta}_2 = 1.1150422, \quad \hat{\beta}_3 = -0.0002717,$$

so the fitted equation is

$$\widehat{\text{Sales}} = 13.1985009 - 0.0546694\,\text{Price} + 1.1150422\,D_{\text{US}} - 0.0002717(\text{Price} \cdot D_{\text{US}}).$$

Equivalently, written by cases:

$$\widehat{\text{Sales}} = \begin{cases} 13.1985009 - 0.0546694\,\text{Price}, & \text{US} = \text{No}, \\ (13.1985009 + 1.1150422) + (-0.0546694 - 0.0002717)\text{Price}, & \text{US} = \text{Yes}. \end{cases}$$

That is,

$$\widehat{\text{Sales}} = \begin{cases} 13.1985009 - 0.0546694\,\text{Price}, & \text{US} = \text{No}, \\ 14.3135431 - 0.0549411\,\text{Price}, & \text{US} = \text{Yes}. \end{cases}$$

Also (I'm not sure if I'm supposed to report on this, but), the interaction term is not statistically significant ($p = 0.985 > \alpha = 0.5$), so there is no evidence (in this training sample) that the slope of Sales versus Price differs between US and non-US stores.

Problem 2 (ii)

```r
# Split training data by US status.
train_us_no  <- dplyr::filter(train, US == 'No')
train_us_yes <- dplyr::filter(train, US == 'Yes')


# Fit separate simple regressions: Sales ~ Price.
```

```r
m_no  <- lm(Sales ~ Price, data = train_us_no)
m_yes <- lm(Sales ~ Price, data = train_us_yes)


# Check both models.
summary(m_no)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = train_us_no)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4395 -1.4399 -0.0463  1.2790  6.1844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.19850    1.19516  11.043  < 2e-16 ***
## Price       -0.05467    0.01032  -5.298 1.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 64 degrees of freedom
## Multiple R-squared:  0.3049, Adjusted R-squared:  0.294
## F-statistic: 28.07 on 1 and 64 DF,  p-value: 1.538e-06
```

```r
summary(m_yes)
```

```
##
## Call:
## lm(formula = Sales ~ Price, data = train_us_yes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5297 -1.6035  0.0052  1.3166  7.0110
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.313543   1.184346  12.086  < 2e-16 ***
## Price       -0.054941   0.009922  -5.537 1.58e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.611 on 133 degrees of freedom
## Multiple R-squared:  0.1874, Adjusted R-squared:  0.1812
## F-statistic: 30.66 on 1 and 133 DF,  p-value: 1.579e-07
```

The fitted equations are as follows:

- **US = No:**

$$\widehat{\text{Sales}} = 13.19850 - 0.05467\,\text{Price}.$$

- **US = Yes:**

$$\widehat{\text{Sales}} = 14.31354 - 0.05494\,\text{Price}.$$

In part (i), the interaction model

$$\widehat{\text{Sales}} = \hat{\beta}_0 + \hat{\beta}_1\,\text{Price} + \hat{\beta}_2\,D_{\text{US}} + \hat{\beta}_3(\text{Price}\cdot D_{\text{US}})$$

implies the two fitted lines

$$\widehat{\text{Sales}} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1\,\text{Price}, & \text{US} = \text{No}, \\ (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3)\,\text{Price}, & \text{US} = \text{Yes}. \end{cases}$$

From the reported estimates in (i), those were

$$\widehat{\text{Sales}} = \begin{cases} 13.1985009 - 0.0546694\,\text{Price}, & \text{US} = \text{No}, \\ 14.3135431 - 0.0549411\,\text{Price}, & \text{US} = \text{Yes}. \end{cases}$$

These match exactly the coefficients obtained by fitting the two separate regressions. In particular, the estimated slopes are very close:

$$\hat{\beta}_{1,\text{No}} = -0.05467 \quad \text{and} \quad \hat{\beta}_{1,\text{Yes}} = -0.05494,$$

so the difference in slopes is

$$\hat{\beta}_{1,\text{Yes}} - \hat{\beta}_{1,\text{No}} = -0.05494 - (-0.05467) = -0.00027,$$

which agrees with the interaction estimate from part (i), $\hat{\beta}_3 = -0.00027$. Thus, there is no meaningful evidence here that the *price effect* differs by US status (consistent with my comment in Problem 2 (i)). By contrast, the intercepts differ more noticeably:

$$\hat{\beta}_{0,\text{No}} = 13.19850 \quad \text{and} \quad \hat{\beta}_{0,\text{Yes}} = 14.31354,$$

so

$$\hat{\beta}_{0,\text{Yes}} - \hat{\beta}_{0,\text{No}} = 14.31354 - 13.19850 = 1.11504.$$

Therefore, for a fixed price, the US=Yes regression line lies about 1.115 thousand units above the US=No line. The main difference between the two groups is a vertical shift, not a change in slope.