413HW2P3

Yanzhuo Cao

2023-06-14

Question a

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

```
##
## 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
```

Question b

[Price: The coefficient for Price represents the expected change in Sales for a one-unit increase in Price, holding all other variables constant. A negative coefficient indicates that an increase in Price is associated with a decrease in Sales, assuming all other factors are constant.]

[Urban: The coefficient for Urban is a binary variable. It represents the expected difference in Sales between an urban location (Urban = 1) and a non-urban location (Urban = 0), holding all other variables constant. A negative coefficient suggests that Sales is lower in urban locations compared to non-urban locations.]

[US: The coefficient for US is also a binary variable. It represents the expected difference in Sales between a product manufactured in the US (US = 1) and a product manufactured outside the US (US = 0), holding all other variables constant. A positive coefficient indicates that Sales are higher for products manufactured in the US.]

Question c

[Sales = β_0 + β_1 * Price + β_2 * UrbanYes + β_3 * USYes + e, or to say Sales = 13.043469 - 0.054459 * Price

- 0.021916 * UrbanYes + 1.200573 * USYes + e]

Question d

[From the results in (a), we can reject the null hypothesis H0: $\beta_j = 0$ for Price and US predictors because their corresponding p-values are less than 0.05. However, we cannot reject the null hypothesis for the Urban predictor as its p-value is not significant.]

Question e

[Based on the significant predictors from part (d), we can fit a smaller model that includes only Price and US as predictors.]

```
model1 <- lm(Sales ~ Price + US, data = Carseats)
summary(model1)
```

```
##
## 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
```

Question f

[The smaller model (Sales ~ Price + US) has an R-squared of 0.2393, indicating that approximately 23.93% of the variance in Sales is explained by the predictors. The Adjusted R-squared is 0.2354, which accounts for the number of predictors and sample size.]

[Comparing the model fit statistics between the full model (from part (a)) and the smaller model, the R-squared and Adjusted R-squared values remain the same, suggesting that removing the Urban predictor did not significantly affect the overall model fit.]

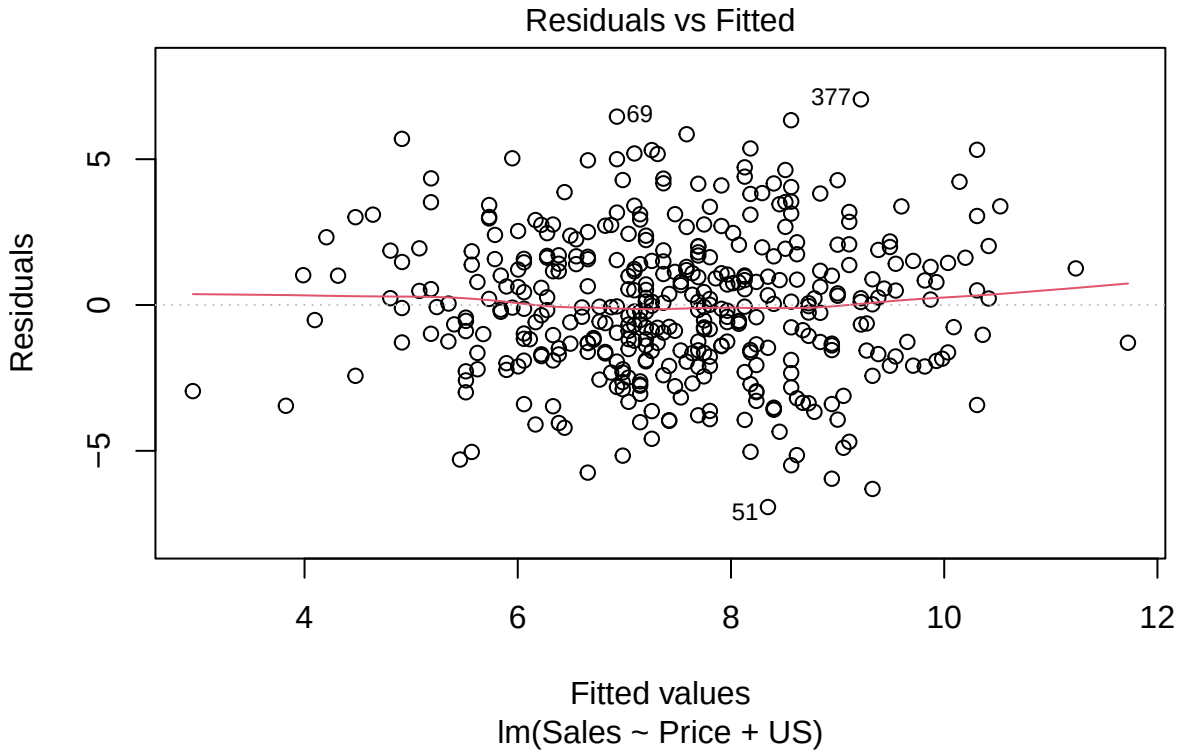
Question g

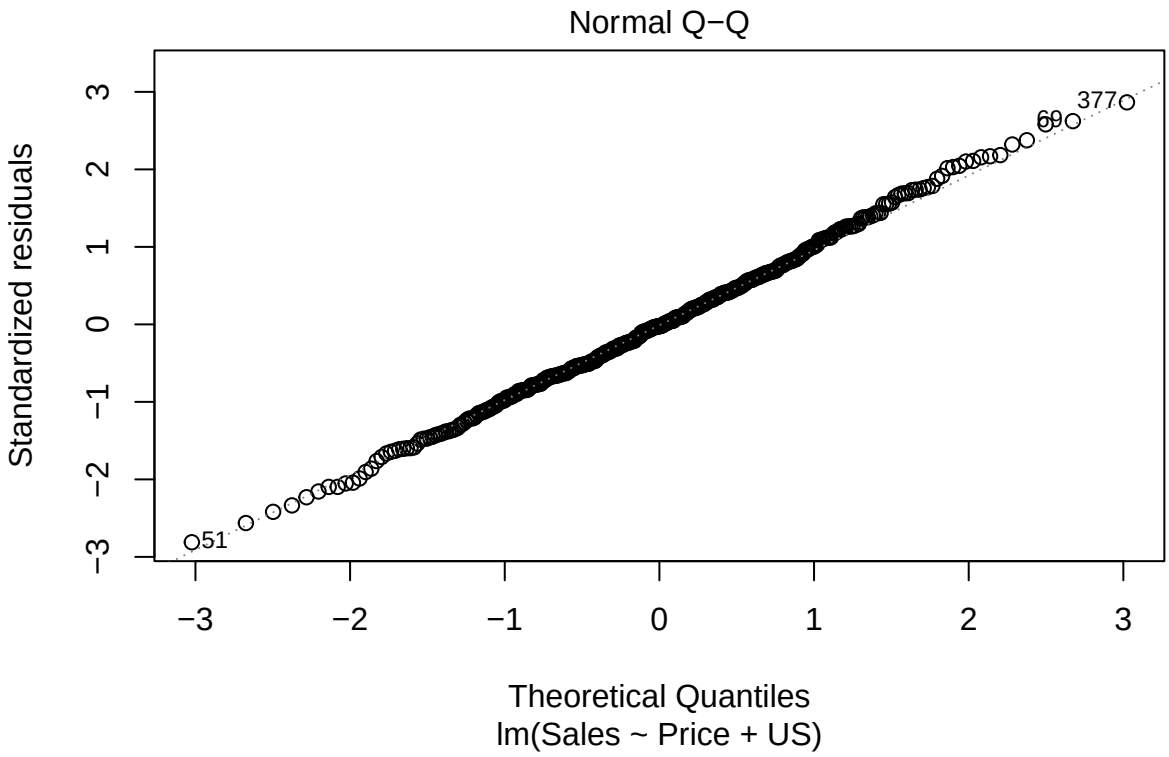
```
confint(model1)
```

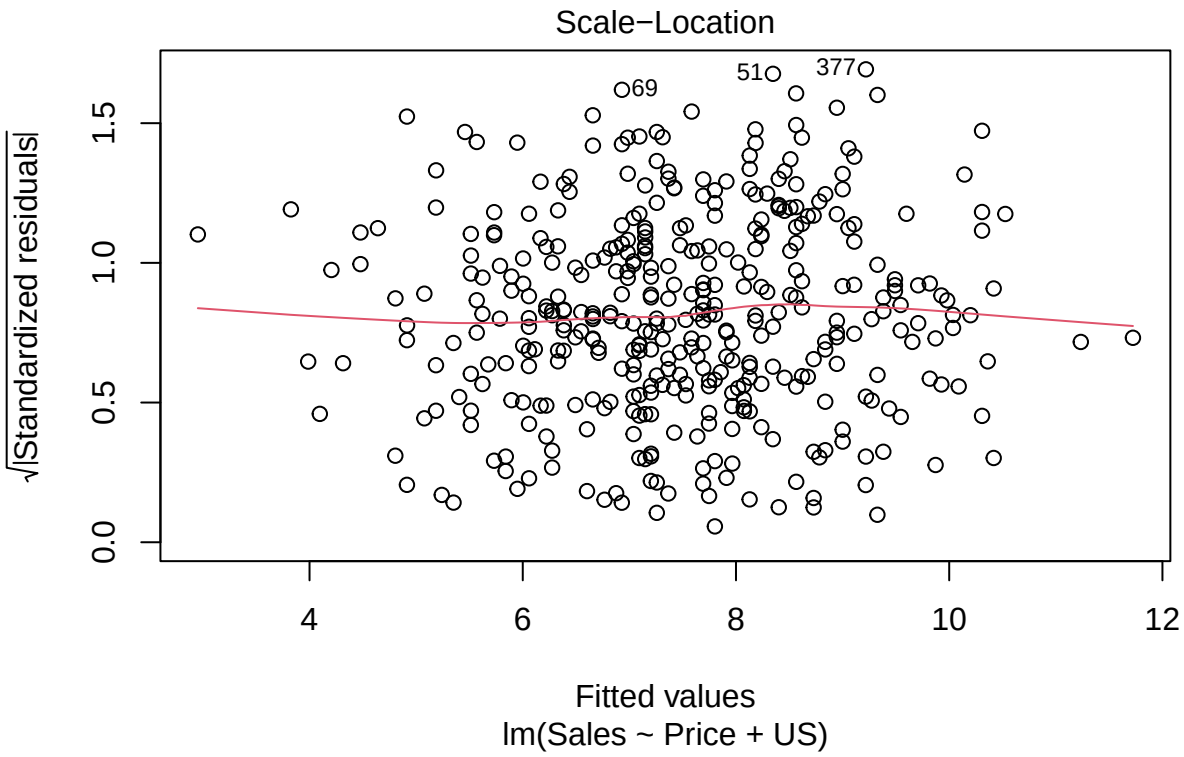
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

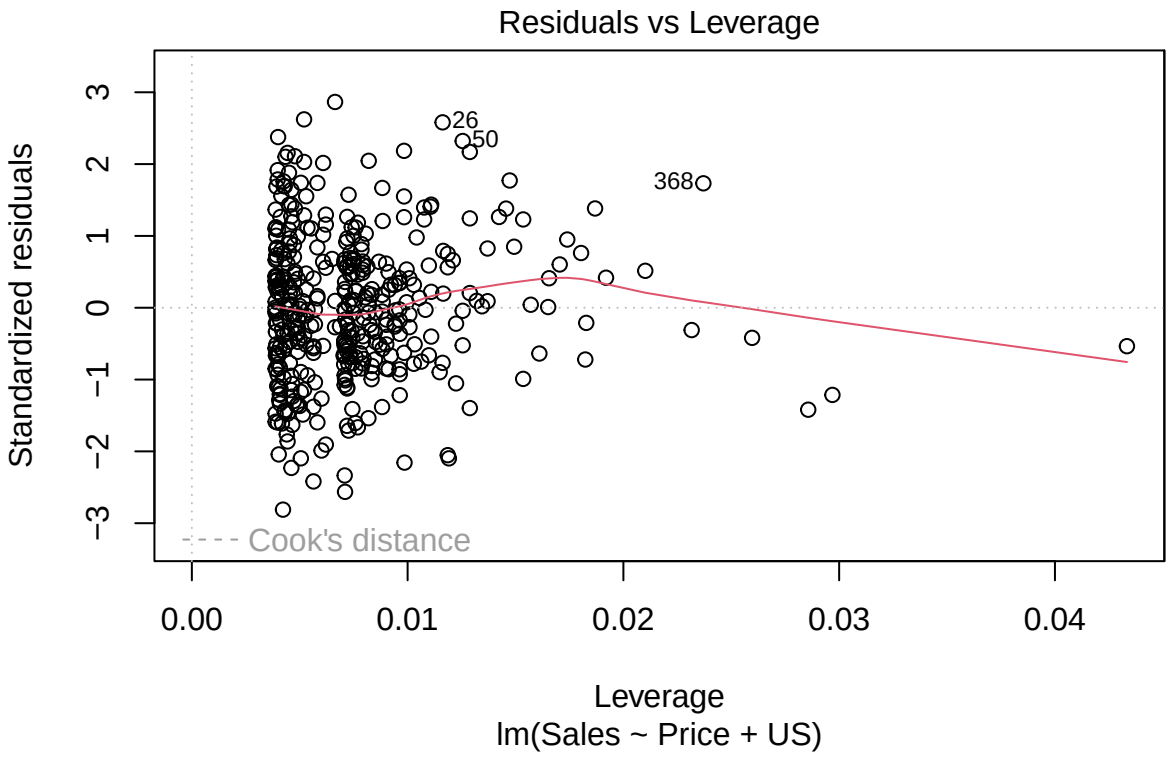
Question h

```
plot(model1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ Price + US)
51
69
377
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
69
51
377
4
6
8
10
12
Fitted values
lm(Sales ~ Price + US)



[We find that the residuals in the residual vs. fitted values plot appear to be randomly distributed, it indicates that the model fits the data well and there are no significant patterns or outliers in the residuals.]

[We find that the points in the normal Q-Q plot roughly fall along a straight line, it suggests that the residuals approximately follow a normal distribution.]

[There is no clear pattern observed in the standardized residuals vs. fitted values plot, it indicates that the assumption of constant variance (homoscedasticity) is likely satisfied.]

[There are high leverage observations are identified in the leverage plot, it suggests that these particular data points have a strong influence on the regression model. These observations exert a significant impact on the estimated coefficients, potentially affecting the overall model fit.It is important to further investigate these high leverage points to determine if they are influential outliers or if they represent genuine characteristics of the data. Additionally, consideration should be given to whether these observations should be retained or removed from the model based on their impact on the model's assumptions and objectives.]

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

Null hypotheses assumes that there is no relationship or effect present in the statistical inference.

① For TV : $p < 0.0001$ ⇒ we will reject the null hypothese.

⇒ there is a significant relationship between TV advertising expenditure and sales

② For ratio : $p < 0.0001$ ⇒ we will reject the null hypothese.

⇒ there is a significant relationship between ratio advertising expenditure and sales

③ For newspaper : $p = 0.8599 > 0.05$ ⇒ we will accept the null hypothese

⇒ there is no significant relationship between newspaper advertising expenditure and sales

3. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

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

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

(a)

$\text{Salary} = 50 + 20\,\text{GPA} + 0.07\,\text{IQ} + 35\,\text{Level} + 0.01 \cdot \text{GPA} \cdot \text{IQ} - 10\,\text{GPA} \cdot \text{Level}$

↳ means that college graduates earn more, on average, than high-school graduates if IQ and GPA are fixed

⇒ ii is correct

(b)

$\text{Salary} = 50 + 20 \cdot 4 + 0.07 \cdot 110 + 35 \cdot 1 + 0.01 \cdot 4 \cdot 110 - 10 \cdot 4 \cdot 1 = 137.1$

⇒ total salary is $137.1 \times 1000 = 137100$ dollars

(c)

False ⇒ ($\hat{\beta}_4 = 0.01$) Although the coefficient for the GPA/IQ interaction term is small, it does not necessarily indicate the absence of an interaction effect. The magnitude of the interaction term coefficient only suggests that its impact on salary is relatively small, but it does not solely determine the presence of an interaction effect. To determine the presence of an interaction effect, statistical hypothesis testing or further analysis is required. The existence of an interaction effect should be assessed based on statistical significance testing or domain knowledge, rather than solely relying on the magnitude of the coefficient. Therefore, a small coefficient for the interaction term does not imply the absence of an interaction effect.