

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 = $\beta_0 + \beta_1 * \text{Price} + \beta_2 * \text{UrbanYes} + \beta_3 * \text{USYes} + e$, 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 $H_0: \beta_j = 0$ 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.]

```
modell1 <- lm(Sales ~ Price + US, data = Carseats)
summary(modell1)

##
## 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.]

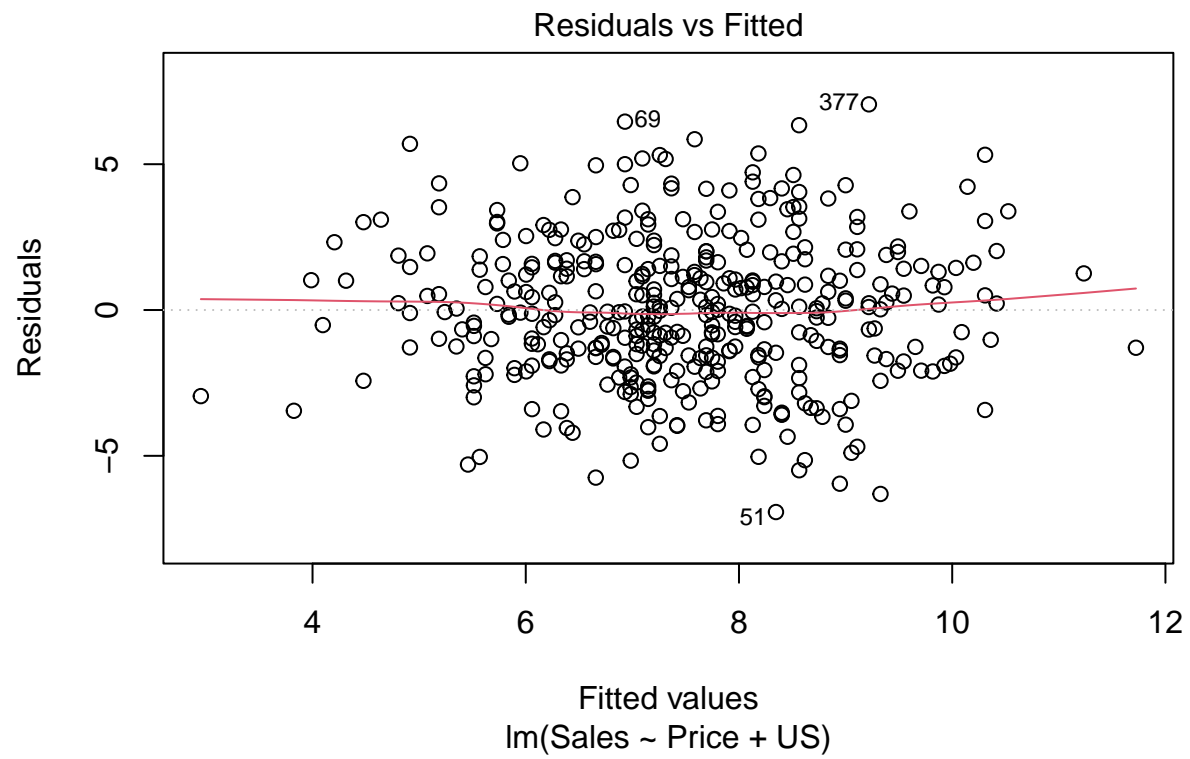
Question g

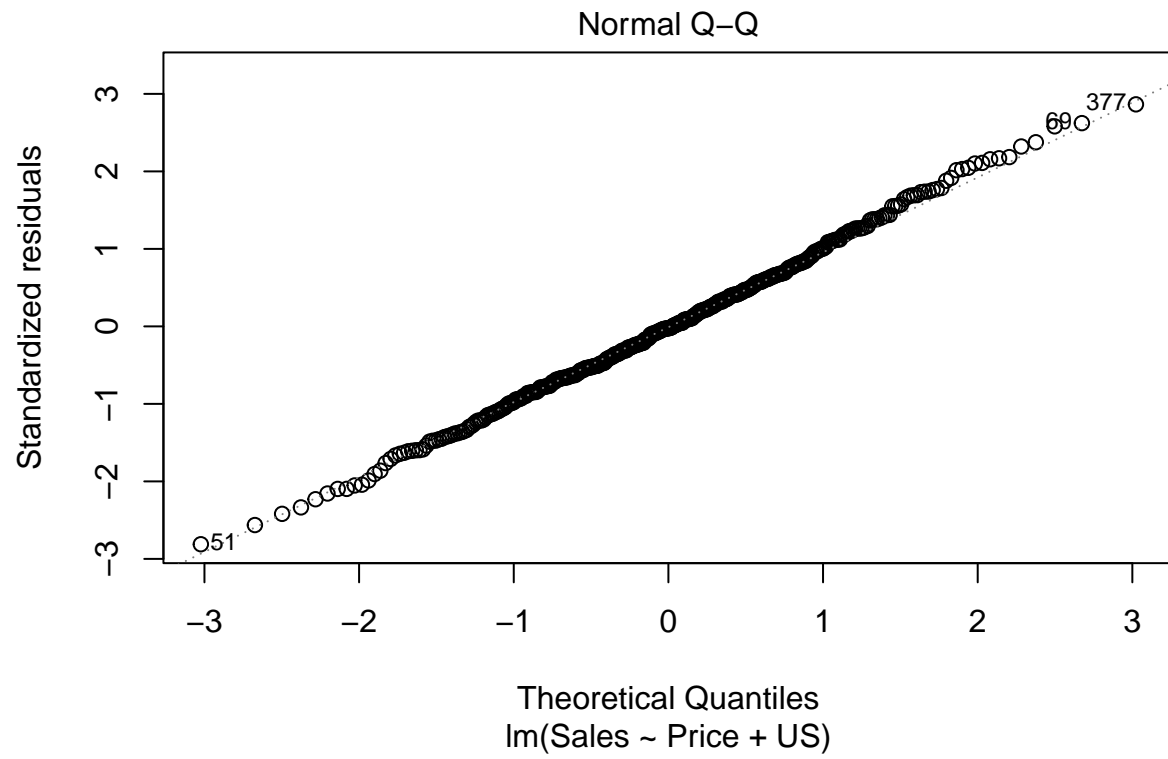
```
confint(modell1)

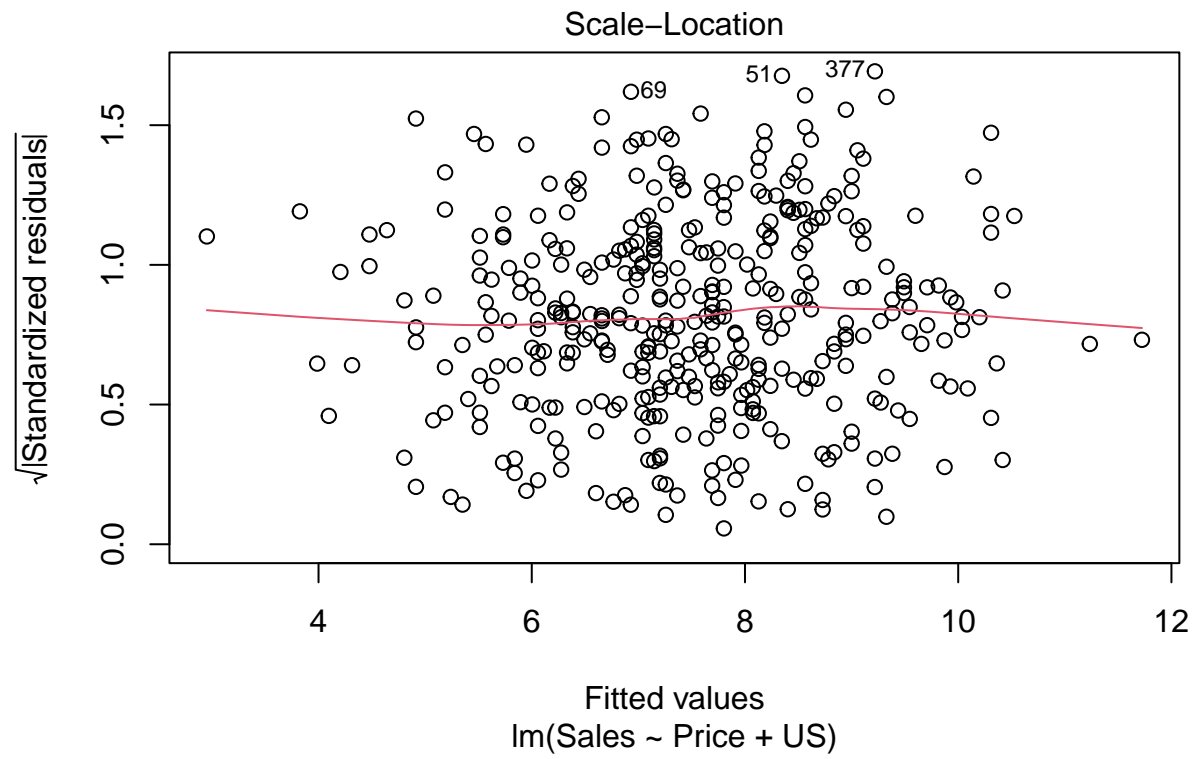
##              2.5 %       97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

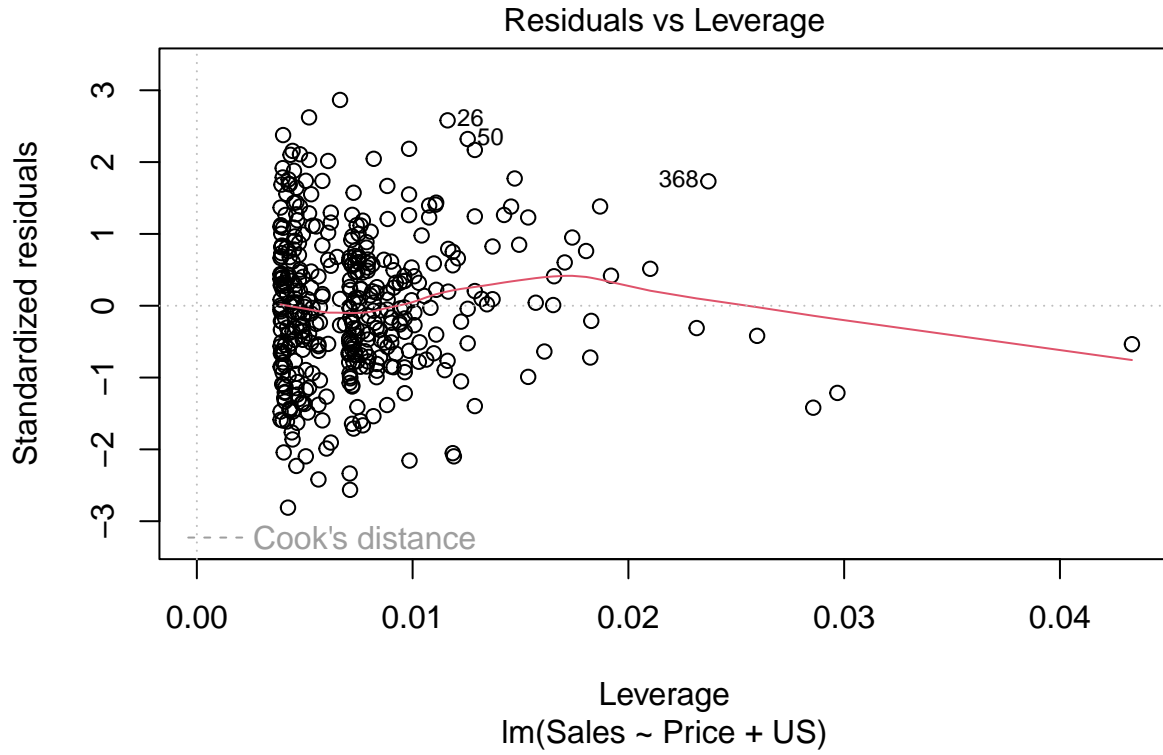
Question h

```
plot(model1)
```









[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.]