# Statistical Method HW6

RE6121011 徐仁瓏

2023-11-10

## 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.

```
data = read.csv('/Users/xurenlong/Desktop/Statistical Methods/Assignments/HW6/Carseats.csv')
model1 = lm(Sales ~ Price + Urban + US, data= data)
summary(model1)
```

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

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

> When all independent variables are zero, the predicted value of `Sales` is 13.043469.
>
> When `Price` increases by one unit, `Sales`, on average, decreases 0.054459 by units.
>
> When `UrbanYes` increases by one unit, `Sales`, on average, decreases 0.021916 by units.
>
> When `USYes` increases by one unit, `Sales`, on average, increases 1.200573 by units.

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

> $$Y = 13.043469 - 0.054459 \cdot X_1 - 0.021916 \cdot X_2 + 1.200573 \cdot X_3$$
>
> Where:
>
> $Y$ is the dependent variable `Sales`
>
> $X_1$ is the independent variable `Price`
>
> $X_2$ represents `Urban`, taking the value 1 for “Yes” and 0 for “No”
>
> $X_3$ represents `US`, taking the value 1 for “Yes” and 0 for “No”

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

> We can reject the null hypothesis for `Price` and `USYes` because the P-value of both predictors are smaller than 0.05.

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

```
model2 = lm(Sales ~ Price + US, data= data)
summary(model2)
```

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

| model | Residual.standard.error | Adjusted.R.squared |
|---|---|---|
| a | 2.472 | 0.2335 |
| e | 2.469 | 0.2354 |

> When comparing two models with the same dependent variable $Y$ but different parameters, we can assess their performance by examining the Residual Standard Error (RSE) and adjusted $R^2$.
>
> A smaller RSE indicates better model fit, while a larger adjusted $R^2$ suggests a more favorable model. According to the table above, it is evident that **Model (e)** outperforms the other model.

## (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.

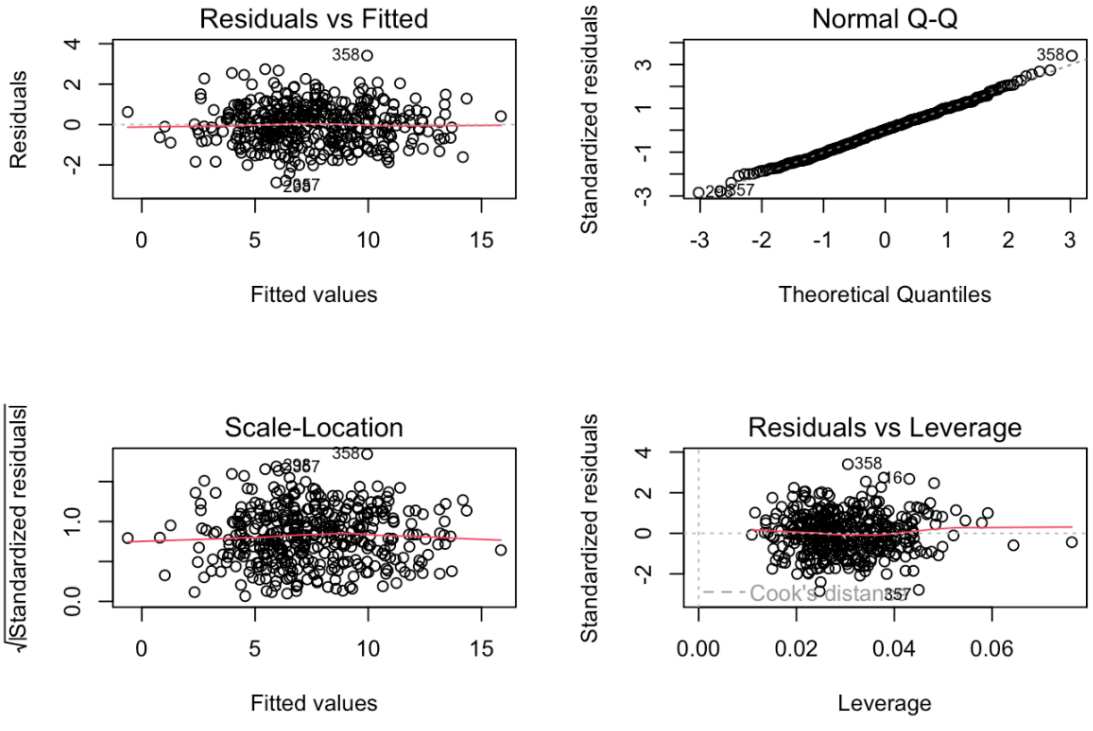
```
model3 = lm(Sales ~ . , data= data)
summary(model3)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = data)
##
## 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 adjusted $R^2$ is 0.8698, which is much larger than model in (a) and (e). Therefore, this model is a better regression model.

## Residual Diagnostics

```
par(mfrow = c(2,2))
plot(model3)
```



> From the first and third plots, it can be observed that there is no apparent pattern in the residuals. Therefore, it seems reasonable to assume homogeneity of variances for this dataset.
>
> In the second plot, although there are a few outliers, the majority of the data points are closely aligned with the reference line, indicating adherence to the normality assumption.
>
> The last plot indicates that there are no points exceeding Cook's distance, suggesting the absence of influential points.
>
> Furthermore, from the plots, it is evident that there are three distinct outliers at observations 298, 357, and 358.

## Normality Assumption

```
shapiro.test(model3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.99758, p-value = 0.8337
```

> Because the p-value is greater than 0.05, it conforms to the normality assumption.

## Homoscedasticity Assumption

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 7.3287, df = 11, p-value = 0.7719
```

> Because the p-value is greater than 0.05, it conforms to the homoscedasticity assumption.

## Independence Assumption

```
dwtest(model3)
```

```
##
##  Durbin-Watson test
##
## data:  model3
## DW = 2.0127, p-value = 0.5509
## alternative hypothesis: true autocorrelation is greater than 0
```

> Because the p-value is greater than 0.05, it conforms to the independence assumption.

---

**2. Suppose we have a data set with five predictors:**

$$X_1 = GPA$$

$$X_2 = IQ$$

$$X_3 = Level(1 for College and 0 for HighSchool)$$

$$X_4 = Interaction between GPA and IQ$$

$$X_5 = Interaction between GPA and Level$$

**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) True of False

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 graduatesprovided that the GPA is high enough.

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

v. Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence ofan interaction effect. Justify your answer.

i. False.

ii. False.

iii.True. The proof is in the following figure.

iv. False.

v. False. The significance of this effect cannot be determined solely by looking at the coefficient.

$\hat{y} = 50 + 20\,GPA + 0.07\,IQ + 35\,level + 0.01\,GPA \times IQ - 10\,GPA \times level$

$\hat{y}_C = 50 + 20\,GPA + 0.07\,IQ + 35 \cdot 1 + 0.01\,GPA \times IQ - 10\,GPA \times 1$

$= 85 + 10\,GPA + 0.07\,IQ + 0.01\,GPA \times IQ$

$\hat{y}_H = 50 + 20\,GPA + 0.07\,IQ + 35 \cdot 0 + 0.01\,GPA \times IQ - 10\,GPA \times 0$

$= 50 + 20\,GPA + 0.07\,IQ + 0.01\,GPA \times IQ$

$\hat{y}_H - \hat{y}_C = -35 + 10\,GPA$

令 $\hat{y}_H - \hat{y}_C > 0 \Rightarrow -35 + 10\,GPA > 0 \Rightarrow GPA > 3.5$

∴當GPA達到3.5以上時，高中生的「平均」起薪才會比大學生高。※

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

```
Salary = 50 + 20*4 + 0.07*110 + 35*1 + 0.01*4*110 - 10*4*1
Salary
```

```
## [1] 137.1
```

$$Salary = 137.1$$