Statistical Method HW5

RE6124019

2023-11-11

```
library(lmtest)
```

Question 1

```
Carseats <- read.csv("C:/Users/wumin/OneDrive/Desktop/SM/Carseats.csv")
```

(a)

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

(b)

When all independent variables are zero, the baseline value for sales is 13.0434.

For each additional unit of Price, sales are estimated to decrease by approximately 0.0545 units.

For each additional unit of UrbanYes, sales are estimated to decrease by approximately 0.0219 units.

For each additional unit of USYes, sales are estimated to increase by approximately 1.2006 units.

(c)

$Y = 13.04346894 - 0.05445885 * X_1 - 0.02191615 * X_2 + 1.20057270 * X_3 + \epsilon$

Where:

Y is the predicted sales value.

X_1 is the price variable.

X_2 is a binary variable indicating whether the observation is in an urban area (1 if yes, 0 if no).

X_3 is a binary variable indicating whether the observation is in the US (1 if yes, 0 if no).

ϵ is the error term.

(d)

We can reject the null hypothesis for Price because the p-value $1.609917e^{-22} < 0.05$.

We can reject the null hypothesis for US because the p-value $4.860245e^{-06} < 0.05$.

(e)

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

For model (a): Adjusted R-squared is 0.2335.

For model (e): Adjusted R-squared is 0.2354.

Both models (a) and (e) have relatively low adjusted R-squared values (0.2335 and 0.2354), indicating that they explain only a small proportion of the variation in the dependent variable.

I suggest considering additional predictors to improve the models' explanatory power.

(g)

```
model_g <- lm(Sales ~ . , data = Carseats)
summary(model_g)
```

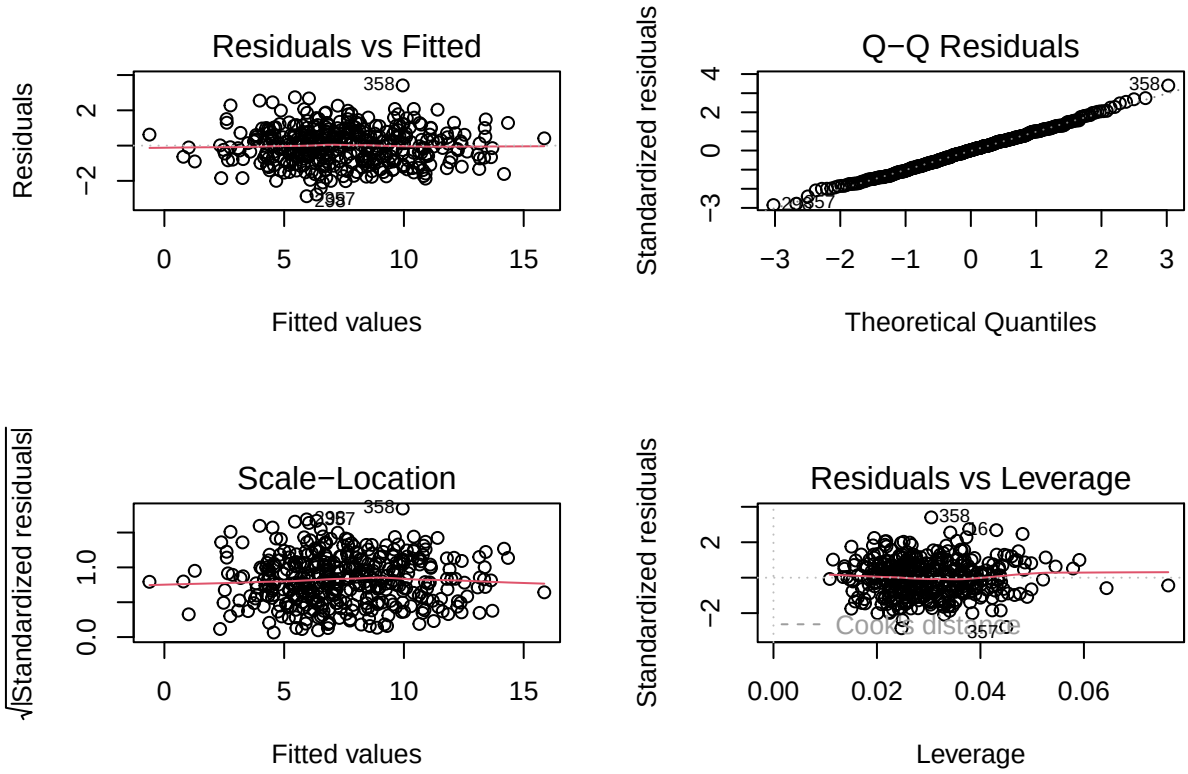
```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***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

The adjusted R-squared: 0.8698, is larger than model_a and model_e.

This model is a better model.

Diagnostic Figures of Residuals

```r
par(mfrow = c(2, 2))
plot(model_g)
```



Residuals vs Fitted:

The residuals are randomly scattered around zero, with no apparent pattern. This meets the model assumption.

Normal Q-Q:

The points follow the diagonal line fairly closely. This supports the assumption that the residuals are approximately normally distributed.

Scale-Location:

The spread is fairly constant. There is no evidence of increasing residual spread as fitted values change.

Residuals vs Leverage:

No high leverage points or outliers are visible.

i. Normal Assumption

```r
shapiro.test(model_g$residuals)
```

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

p-valu > 0.05, it meets the Normal Assumption

ii. Independent Assumption

```
dwtest(model_g)
```

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

p-valu > 0.05, it meets the Independent Assumption

iii. Homoscedasticity Assumption

```
bptest(model_g)
```

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

p-valu > 0.05, it meets the Homoscedasticity Assumption

Question 2

(a)

$\hat{Y} = 50 + 20 * GPA + 0.07 * IQ + 35 * Level + 0.01 * GPA * IQ - 10 * GPA * Level$

$\hat{Y}_{College} = 50 + 20 * GPA + 0.07 * IQ + 35 * 1 + 0.01 * GPA * IQ - 10 * GPA * 1$

$\hat{Y}_{HighSchool} = 50 + 20 * GPA + 0.07 * IQ + 35 * 0 + 0.01 * GPA * IQ - 10 * GPA * 0$

For a fixed value of IQ and GPA

$\hat{Y}_{College} = 35 * 1 - 10 * GPA * 1$

$\hat{Y}_{HighSchool} = 35 * 0 - 10 * GPA * 0$

$\hat{Y}_{College} - \hat{Y}_{HighSchool} = 35 - 10 * GPA$

$Let \hat{Y}_{College} - \hat{Y}_{HighSchool} > 0; GPA > 3.5$

i False.

The answer may not necessarily be definitive; we need to consider the factor "GPA".

ii False.

The answer may not necessarily be definitive; we need to consider the factor "GPA".

iii True.

If the GPA is greater than 3.5, high school graduates earn more, on average, than college graduates.

iv False.

If the GPA is greater than 3.5, college graduates earn less, on average, than high school graduates.

v False.

If GPA*IQ is a very large value, it will still have an impact on the outcome.

(b)

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

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

Salary = 137.1