STAT 4130: Homework (Template)

Your name

2023-06-08

```r
knitr::opts_chunk$set(echo = TRUE)
# if you are using libraries, it's good practice to load them here
library(ISLR)
```

Question 1

The p-value of TV and radio are both < 0.0001. So the factor “sales” is positively correlated with the factor “TV” and “radio”. The p-value of “newspaper” is 0.8599, which is a relatively high value. So there is no strong evidence that the factor “sales” is negatively correlated with the factor “newspaper”.

Question 2

(a) Y = 50 + 20 * GPA + 0.07 * IQ + 35 * Level + 0.01 * GPA * IQ - 10 * GPA * Level I choose iii. Firstly, i and ii does not provide the information about GPA. If GPA > 3.5, 10 * 3.5 * Level > 35 * Level, which make Y negatively correlated with Level. For iii and iv, the additional condition is that GPA is high enough, in this case, Y is negatively correlated with Level, which makes high school graduates earning more than college students.

(b) For a college student, Level = 1, GPA = 4.0, IQ = 110: After plugging in the formula: Y = 50 + 20 * 4.0 + 0.07 * 110 + 35 * 1 + 0.01 * 4.0 * 110 - 10 * 4.0 * 1 = 137.1.

(c) False. We cannot simply reach the conclusion that there is very little evidence of an interaction effect. We need to do a F-test to find the necessity of the interaction term based on the data.

Question 3

```r
# Your Code
# (a)
lm1 = lm(formula = Sales ~ Price + Urban + US, data = Carseats)
# (d)
lm2 = lm(formula = Sales ~ Price + Urban, data = Carseats)
lm3 = lm(formula = Sales ~ Price + US, data = Carseats)
lm4 = lm(formula = Sales ~ Urban + US, data = Carseats)
anova(lm1, lm2)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Price + Urban
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1    396 2420.8
## 2    397 2552.2 -1   -131.31 21.48 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm1, lm3)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Price + US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    396 2420.8
## 2    397 2420.9 -1  -0.03979 0.0065 0.9357
```

```
anova(lm1, lm4)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Urban + US
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    396 2420.8
## 2    397 3080.7 -1   -659.84 107.94 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# (f)
summary(lm1)
```

```
##
## 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
```

```
summary(lm3)
```

```
##
## 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
```

```r
# (g)
confint(lm1)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.76359670 14.32334118
## Price       -0.06476419 -0.04415351
## UrbanYes    -0.55597316  0.51214085
## USYes        0.69130419  1.70984121
```

```r
confint(lm3)
```

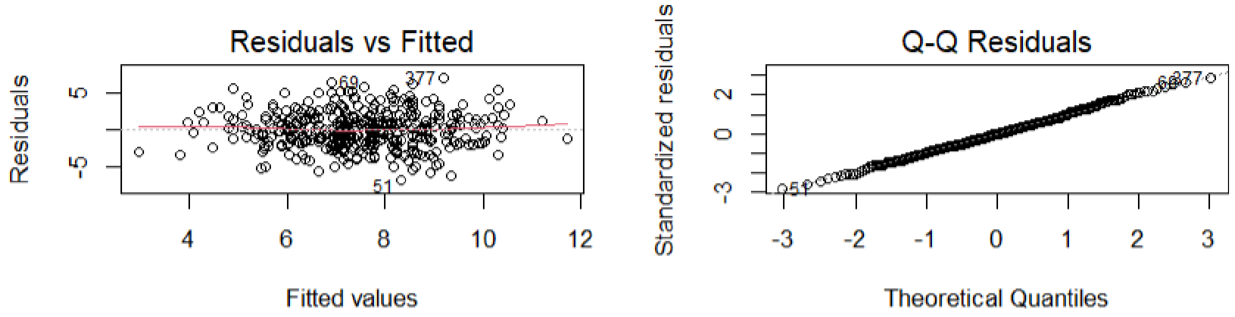
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

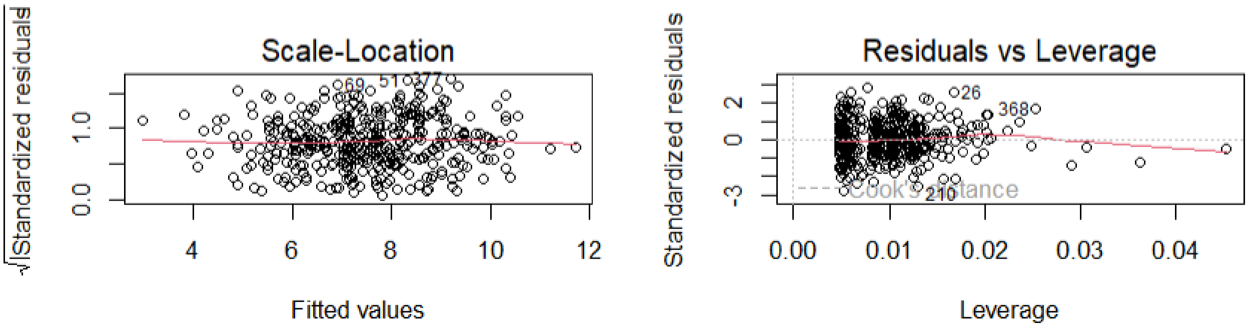
```r
# (h)
# for lm1
par(mfrow = c(2,2))
plot(lm1)
```

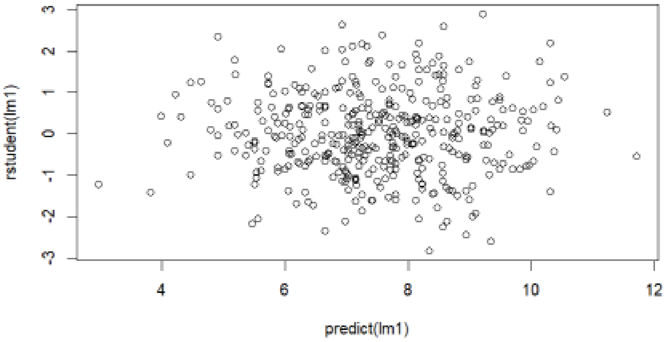
```r
par(mfrow = c(1,1))
plot(predict(lm1),rstudent(lm1))
```

```r
which.max(hatvalues(lm1))
```

```
## 43
## 43
```





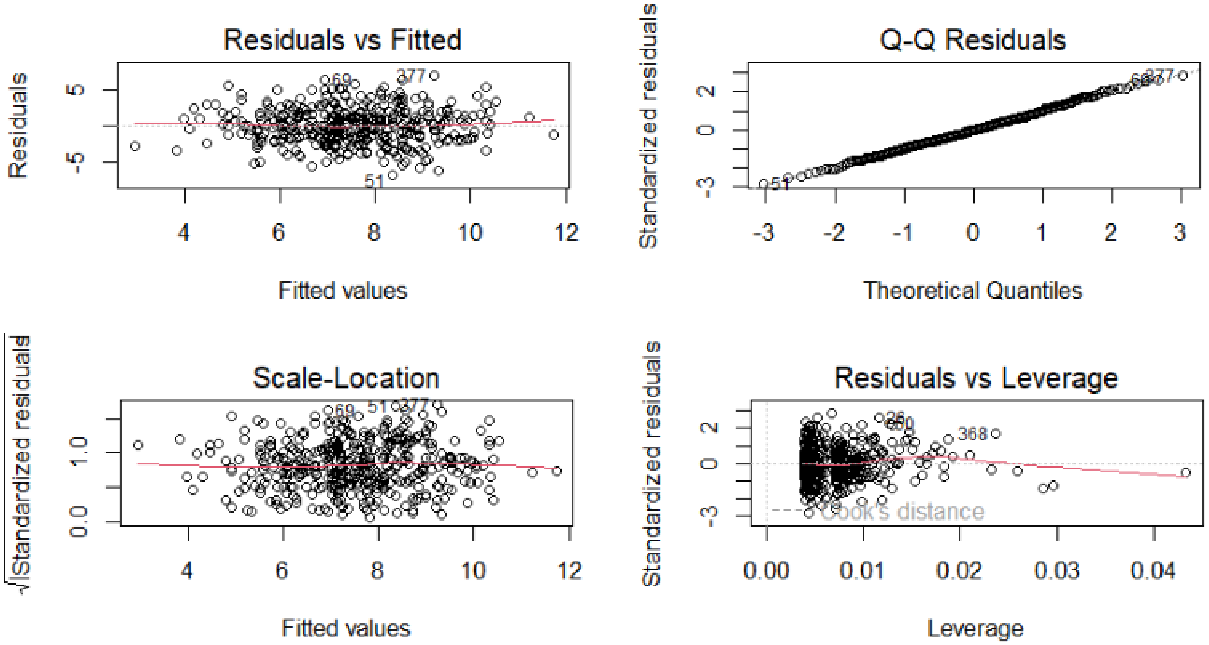


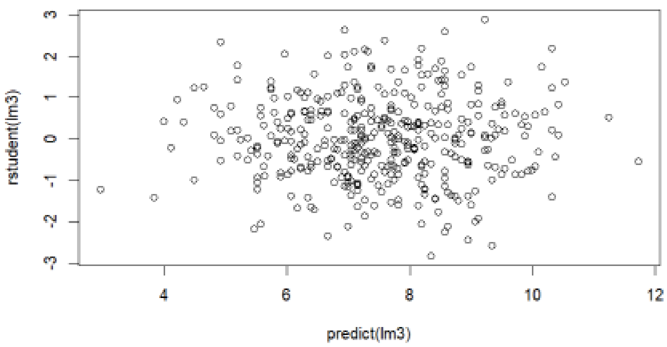
```
# for lm3
par(mfrow = c(2,2))
plot(lm3)
```

```
par(mfrow = c(1,1))
plot(predict(lm3),rstudent(lm3))
```

```
which.max(hatvalues(lm3))
```

```
## 43
## 43
```





3b The intercept of this model is 13.04347, which means Sales = 13.04347 if Price = Urban = US = 0. The coefficient of Price is -0.05446, which means predicted Sales decreases by 0.05446 for each extra Price.(Sales is negatively correlated with Price) The coefficient of UrbanYes is -0.02192, which means the Sales will differ by 0.02192, where urban area owns a lower sales. Similarly, the coefficient of USYes is 1.20057, which means the Sales will differ by 1.20057, where US area owns a higher sales.

3c Sales = 13.04347 - 0.05446 * Price - 0.02192 * UrbanYes + 1.20057 * USYes.

3d For the predictor "USYes" and "Price" we can reject the null hypothesis $\beta_j = 0$. We conducted three anova table for the full model lm1 and reduced model lm2, lm3 and lm4. The p-value of them are 4.86e-06, 0.9357 and $<$ 2.2e-16 for the test of "USYes", "UrbanYes" and "Price" respectively. So we can reach the conclusion that reduced model can not be applied for "USYes" and "Price". In other words, we can reject the null hypothesis $\beta_j = 0$ for the predictor "USYes" and "Price".

3e As there is no evidence that "UrbanYes" predictor is necessary in the linear model, the linear model lm3 is the reduced model that is required for this question.

3f Both of the models fit the data well as their p-values are both $<$ 2.2e-16 according to summary(lm1) and summary(lm3).

3g According to confint(lm1) and confint(lm3), we can reach the two-sided 95% confidence interval of all of the coefficients:
In lm1: 2.5 % 97.5%
(Intercept) 11.76359670 14.32334118
Price -0.06476419 -0.04415351
UrbanYes -0.55597316 0.51214085
USYes 0.69130419 1.70984121
In lm3: 2.5% 97.5 %
(Intercept) 11.79032020 14.27126531
Price -0.06475984 -0.04419543
USYes 0.69151957 1.70776632