Fall 2022 STAT 706 HW 5

Reina Li

Chapter 5: 5.4.1, 5.4.2, 5.4.3

5.4.1

```
#Load data
overdue <- read.table("overdue.txt", header = TRUE)

#Dummy variable
#First 48 accounts are residential = 0
#Second 48 accounts are commercial = 1
overdue$TYPE <- c(rep(0,48), rep(1,48))

#Fit model
model1 <- lm(LATE ~ BILL + TYPE + TYPE:BILL,
             data = overdue)
summary(model1)
```

```
##
## Call:
## lm(formula = LATE ~ BILL + TYPE + TYPE:BILL, data = overdue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1211  -2.2163   0.0974   1.9556   8.6995
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.209624   1.198504   1.844   0.0685 .
## BILL         0.165683   0.006285  26.362   <2e-16 ***
## TYPE        99.548561   1.694940  58.733   <2e-16 ***
## BILL:TYPE   -0.356644   0.008888 -40.125   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.371 on 92 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9796
## F-statistic:  1524 on 3 and 92 DF,  p-value: < 2.2e-16
```

The model is:

$Y = \beta_0 + \beta_1 x + \beta_2 C + \beta_3 C * x + e$, where Y = number of days the payment is overdue; x = the amount of the overdue bill in dollars; and C is a dummy variable which is 0 when the account is residential and 1 when the account is commercial.

$Y = \beta_0 + \beta_1 x + e$ when $C = 0$

$Y = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x + e$ when $C = 1$

So for residential accounts (i.e., $C = 0$) the model predicts: $Days\ Overdue = 2.209624 + 0.165683 * Overdue\ Amount$

and for commercial accounts (i.e., $C = 1$) the model predicts: $Days\ Overdue = 101.758185 - 0.190961 * Overdue\ Amount$

5.4.2

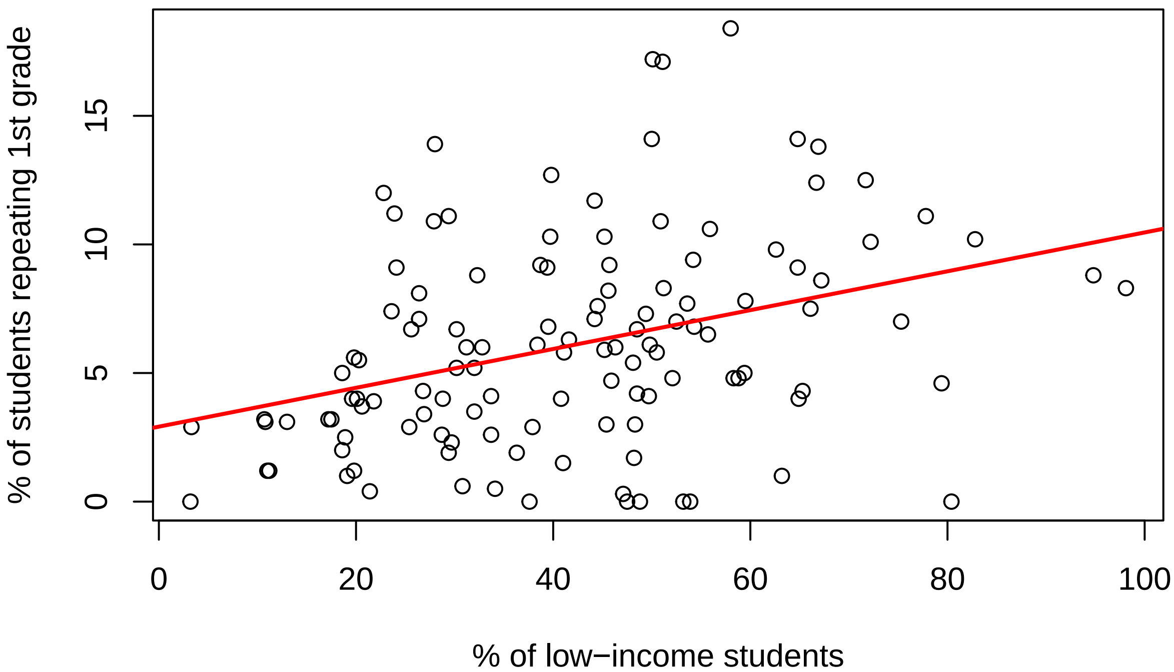
```
#Load data
HoustonChronicle <- read.csv("HoustonChronicle.csv", header = TRUE)

#Dummy variable
#When Year is 2004-2005 = 0
#When Year is 1994-1995 = 1
HoustonChronicle$Dummy <- ifelse(HoustonChronicle$Year == 2004, 0, 1)
```

5.4.2 (a)

```
#Fit model
model2a <- lm(X.Repeating.1st.Grade ~ X.Low.income.students,
              data = HoustonChronicle)
summary(model2a)
```

```
##
## Call:
## lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students, data = HoustonChronicle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9845 -2.5072 -0.4184  1.8505 11.1067
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.91419    0.83836   3.476 0.000709 ***
## X.Low.income.students  0.07550    0.01823   4.141 6.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.821 on 120 degrees of freedom
## Multiple R-squared:  0.125,  Adjusted R-squared:  0.1177
## F-statistic: 17.14 on 1 and 120 DF,  p-value: 6.472e-05
```

```
attach(HoustonChronicle)
plot(X.Low.income.students, X.Repeating.1st.Grade,
     xlab = "% of low-income students", ylab = "% of students repeating 1st grade")
abline(coef(model2a), lwd = 2, col = c("red"))
```




```
detach(HoustonChronicle)
```

Yes, an increase in the percentage of low-income students is associated with an increase in the percentage of students repeating first grade.

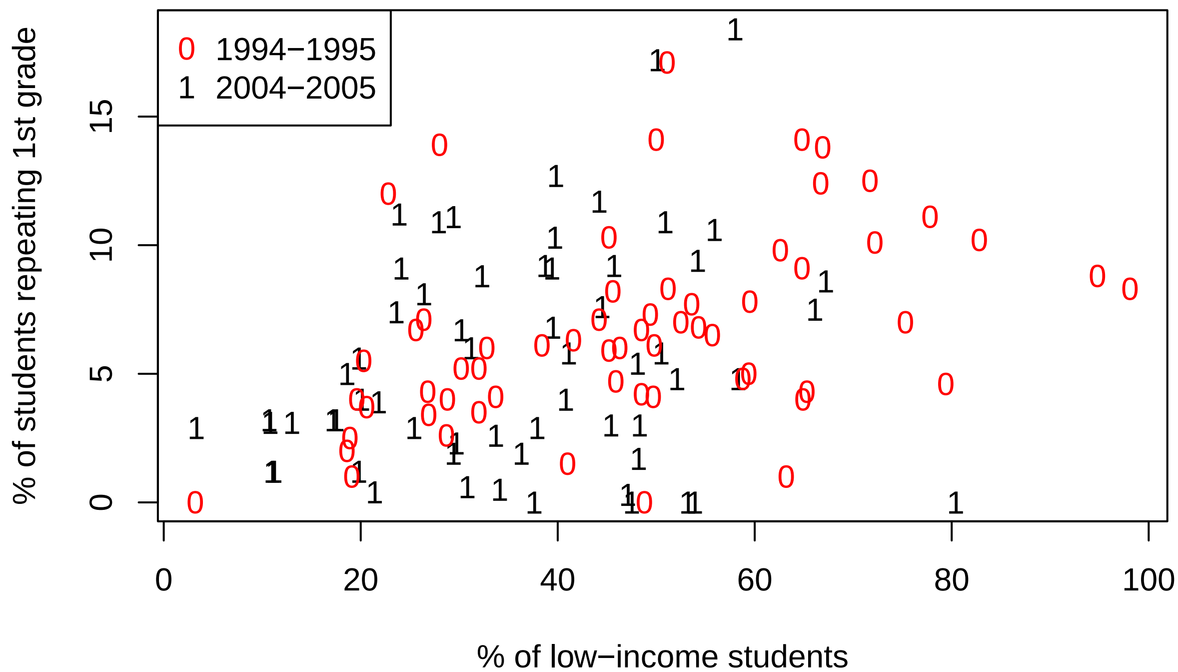
5.4.2 (b)

```
#Fit model
model2b <- lm(X.Repeating.1st.Grade ~ X.Low.income.students + Dummy + Dummy:X.Low.income.students,
             data = HoustonChronicle)
summary(model2b)
```

```
##
## Call:
## lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students +
##     Dummy + Dummy:X.Low.income.students, data = HoustonChronicle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1606 -2.6121 -0.5576  1.7495 11.6014
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                    2.88238    1.26671   2.275  0.02468 *
## X.Low.income.students          0.07984    0.02455   3.253  0.00149 **
## Dummy                          0.38956    1.76109   0.221  0.82532
## X.Low.income.students:Dummy -0.01903    0.03949  -0.482  0.63066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 118 degrees of freedom
## Multiple R-squared:  0.1288, Adjusted R-squared:  0.1066
## F-statistic: 5.813 on 3 and 118 DF,  p-value: 0.0009689
```

```r
attach(HoustonChronicle)
plot(X.Low.income.students[Dummy == 1], X.Repeating.1st.Grade[Dummy == 1],
     pch = c("1"), col = c("black"),
     xlab = "% of low-income students", ylab = "% of students repeating 1st grade",
     xlim = c(min(X.Low.income.students), max(X.Low.income.students)),
     ylim = c(min(X.Repeating.1st.Grade), max(X.Repeating.1st.Grade)))
points(X.Low.income.students[Dummy == 0], X.Repeating.1st.Grade[Dummy == 0],
       pch = c("0"), col = c("red"))
legend('topleft', legend = c("1994-1995", "2004-2005"),
       col = c("red", "black"), pch = c("0", "1"))
```




```r
detach(HoustonChronicle)
```

No, there has not been an increase in the percentage of students repeating first grade between 1994-1995 and 2004-2005.

5.4.2 (c)

No, any association between the percentage of students repeating first grade and the percentage of low-income students does not differ between 1994-1995 and 2004-2005.

5.4.3 (a)

Looking at the Regression Output from R on page 149, the p-value is 0.01203, which is less than 0.05. Therefore, the coefficient of the interaction term in model (5.10) is statistically significant at the 95% level.

5.4.3 (b) (i)

$Quality = \beta_0 + \beta_1 * End\ of\ Harvest + \beta_2 * Rain + \beta_3 * End\ of\ Harvest * Rain + e$

No unwanted rain at harvest = 0

$Quality = 5.16122 - 0.03145 * End\ of\ Harvest + 1.78670 * 0 - 0.08314 * End\ of\ Harvest * 0 + e$

$Quality = 5.16122 - 0.03145 * End\ of\ Harvest + e$

$\frac{\partial Quality}{\partial End\ of\ Harvest} = -0.03145$

$\frac{\partial End\ of\ Harvest}{\partial Quality} = \frac{1}{-0.03145}$

```
1/abs(-0.03145)
```

```
## [1] 31.7965
```

The number of days of delay to the end of harvest it takes to decrease the quality rating by 1 point when there is no unwanted rain at harvest is 32 days.

5.4.3 (b) (ii)

$Quality = \beta_0 + \beta_1 * End\ of\ Harvest + \beta_2 * Rain + \beta_3 * End\ of\ Harvest * Rain + e$

Some unwanted rain at harvest = 1

$Quality = 5.16122 - 0.03145 * End\ of\ Harvest + 1.78670 * 1 - 0.08314 * End\ of\ Harvest * 1 + e$

$Quality = 5.16122 - 0.03145 * End\ of\ Harvest + 1.78670 - 0.08314 * End\ of\ Harvest$

$Quality = 6.94792 - 0.11459 * End\ of\ Harvest + e$

$\frac{\partial Quality}{\partial End\ of\ Harvest} = -0.11459$

$\frac{\partial End\ of\ Harvest}{\partial Quality} = \frac{1}{-0.11459}$

```
1/abs(-0.11459)
```

```
## [1] 8.726765
```

The number of days of delay to the end of harvest it takes to decrease the quality rating by 1 point when there is some unwanted rain at harvest is 9 days.