HW4

Getong Zhong

2022-11-03

4.2 Exercises

Problem 1

```r
setwd("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126")
salary <- read.table("ProfessorSalaries.txt", header = TRUE)
attach(salary)
model <- lm(ThirdQuartile ~ Experience)
wt <- 1/lm(abs(model$residuals) ~ model$fitted.values)$fitted.values^2
wt_model <- lm(ThirdQuartile ~ Experience, weights = wt)
predict(wt_model, data.frame(Experience = 6))
```

```
##        1
## 112752.5
```

Problem 2

Since intercept is 0 then weighted least square estimate of intercept is also 0, therefore

$$\hat{\beta_0}w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} - \beta_1 w * \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = 0$$

$$\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} = \hat{\beta_1}w * \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$\hat{\beta_1}w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i x_i}$$

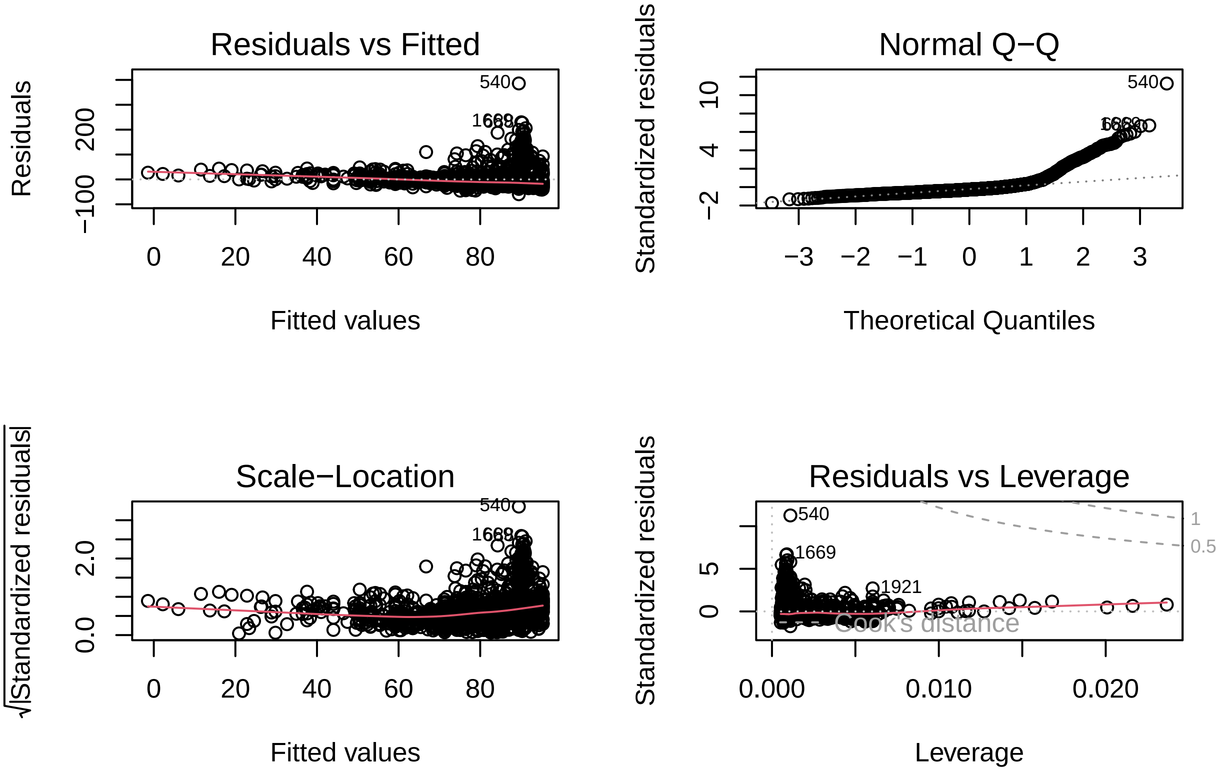
if we take weight as 1/xi then the equation becomes,

$$\hat{\beta_1}w = \frac{\sum_{i=1}^{n} y_i/x_i}{n}$$

Problem 3

(a) It shows that there is clearly heteroskedasticity in this model so we need to use weighted least squares to fit model. ni is valid weights since the response variable Y is calculating the median of a group of variables, we can account the n (simple size of house sold in that year) as the weights.

```
Houston <- read.table("HoustonRealEstate.txt", header = TRUE)
attach(Houston)
model3 <- lm(Yi ~ x1i + x2i)
par(mfrow=c(2,2))
plot(model3)
```



(b) The model (4.6) is not valid because it failed the assumption of equal or similar variances in different groups being compared.

(c) I will try a log transformation to the model since the residuals plot shows it might have a potential log relationship between the predictor and response variable.

5.4 Exercises

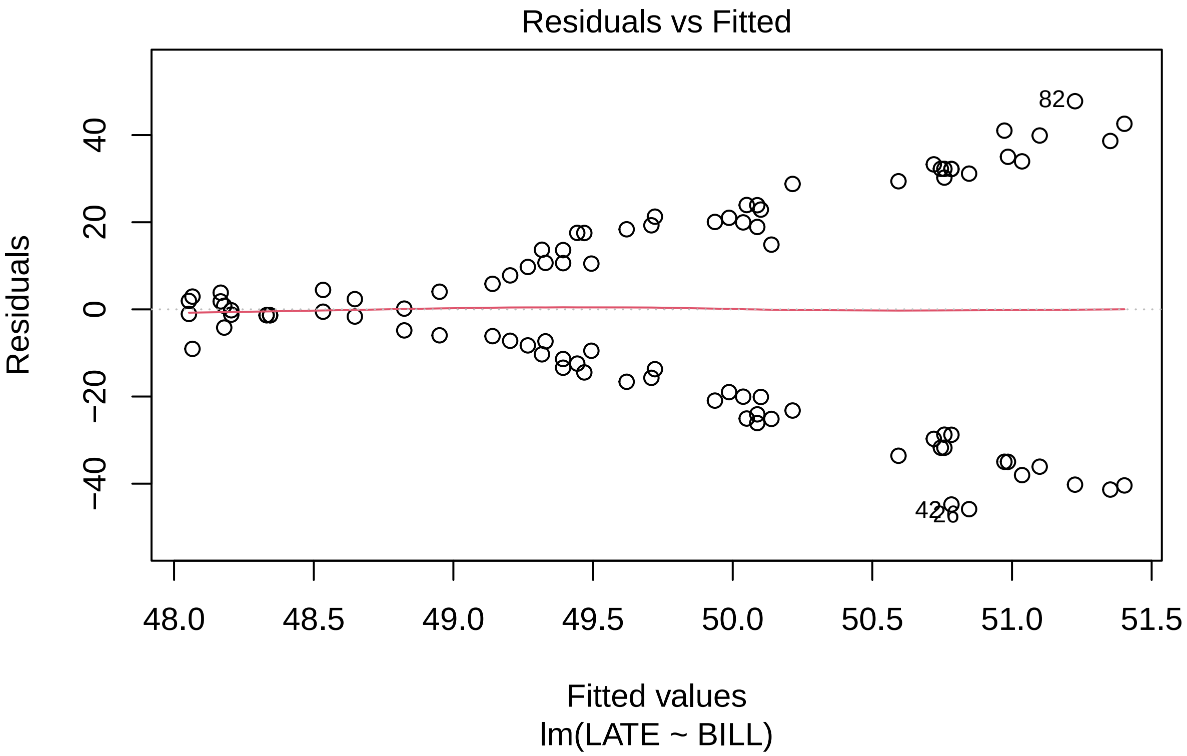
Problem 1

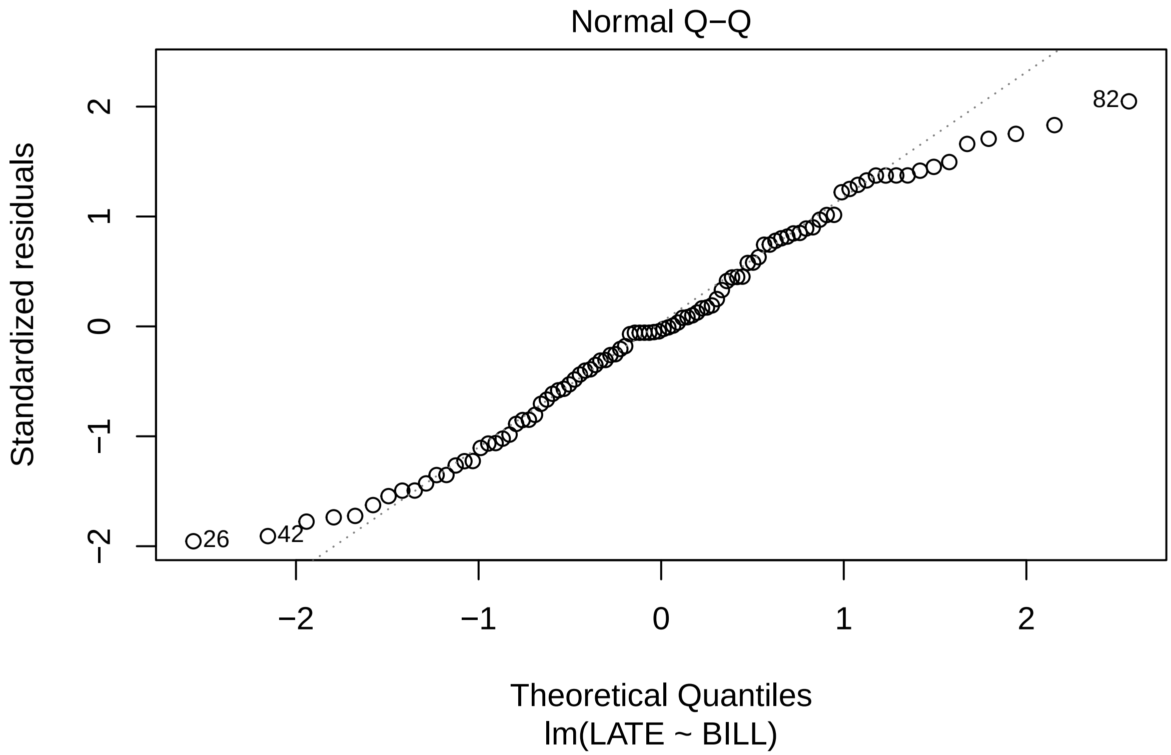
We first just put them into a least square linear regression model, from the summary of the model we can see that the predictor BILL is not statistical significant, therefore some adjustment of the model is necessary.

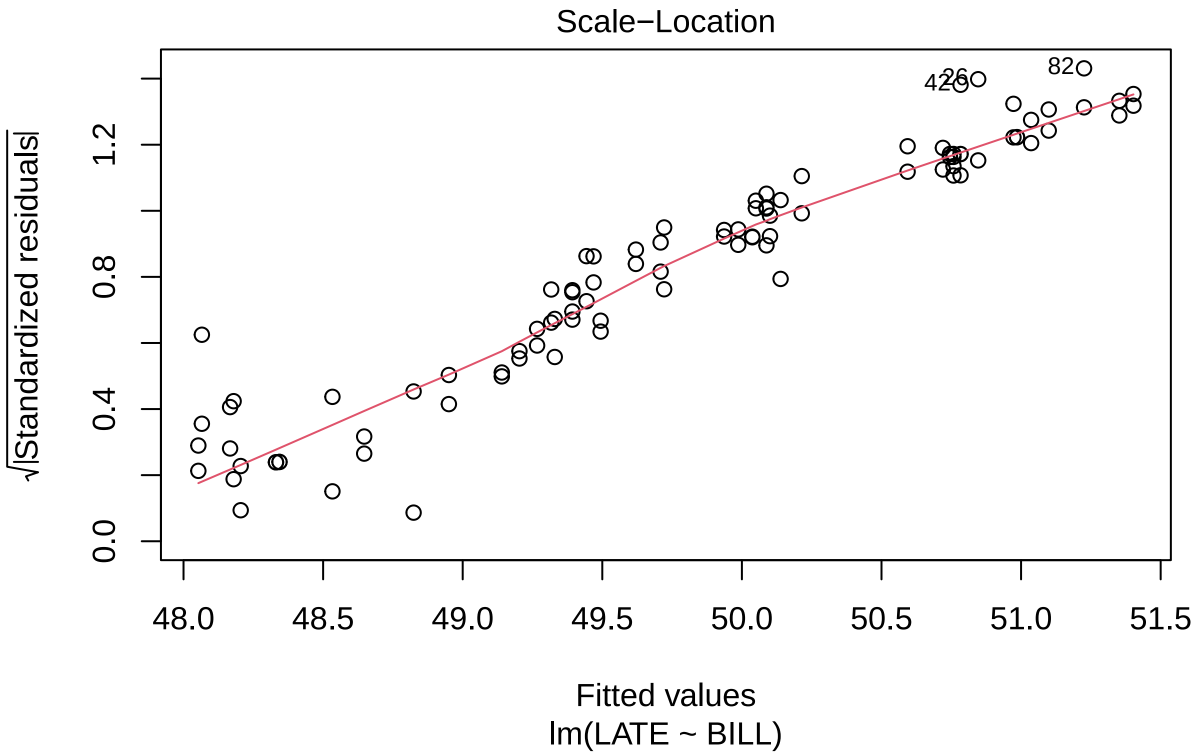
```
overdue <- read.table("overdue.txt", header = TRUE)
attach(overdue)
overdue_model <- lm(LATE ~ BILL)
summary(overdue_model)
```

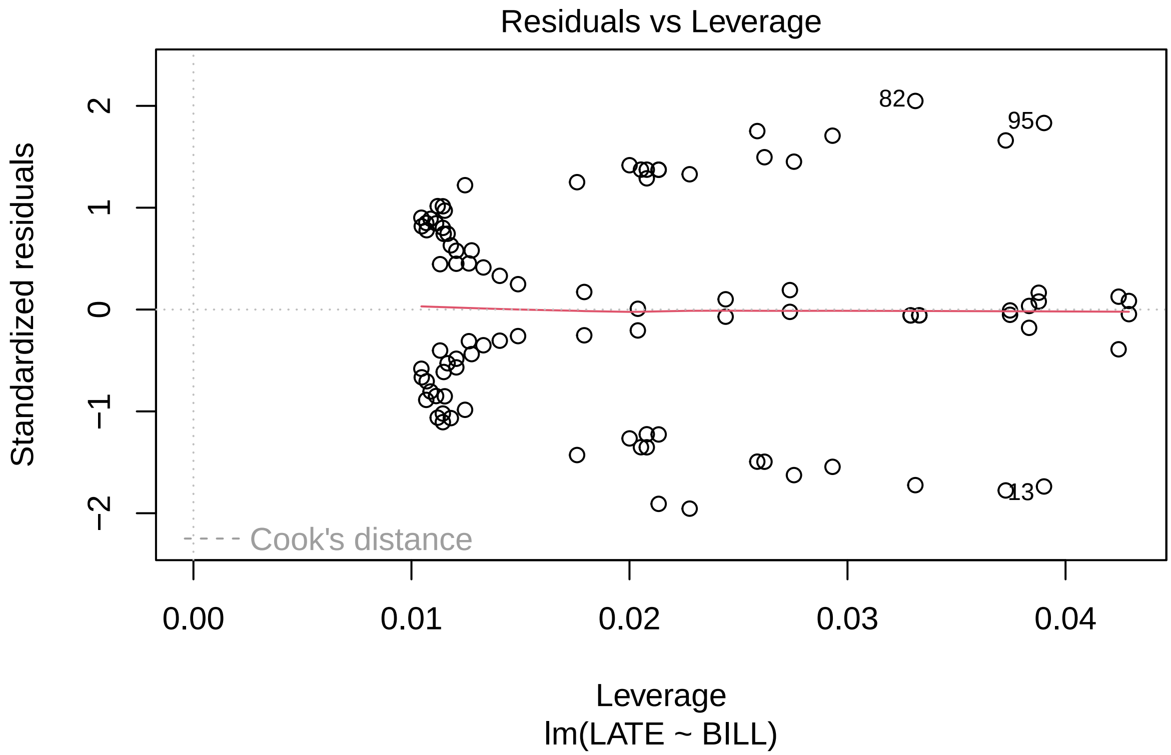
```
##
## Call:
## lm(formula = LATE ~ BILL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.846 -17.212  -0.793  19.007  47.774
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.98390    5.96405   8.716 9.84e-14 ***
## BILL        -0.01264    0.03128  -0.404    0.687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.72 on 94 degrees of freedom
## Multiple R-squared:  0.001734,   Adjusted R-squared:  -0.008885
## F-statistic: 0.1633 on 1 and 94 DF,  p-value: 0.687
```

```
plot(overdue_model)
```




Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(LATE ~ BILL)
26
42
82


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
48.0
48.5
49.0
49.5
50.0
50.5
51.0
51.5
Fitted values
lm(LATE ~ BILL)
42
26
82



From the plots we can clearly observe that there is two types of data which are almost symmetric, therefore we are going to assign the TYPE to each BILL and add the categorical variable TYPE into the model

```
overdue$TYPE <- c(rep(1,48), rep(2,48))
attach(overdue)
```

```
## The following objects are masked from overdue (pos = 3):
##
##     BILL, LATE
```

```
overdue_model2 <- lm(LATE ~ BILL + TYPE)
summary(overdue_model2)
```

```
##
## Call:
## lm(formula = LATE ~ BILL + TYPE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7637 -11.4760   0.4037  12.4812  29.0765
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.10985    5.71330  -0.719    0.474
## BILL        -0.01264    0.01901  -0.665    0.508
## TYPE        37.39583    2.94375  12.703   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.42 on 93 degrees of freedom
## Multiple R-squared:  0.635,  Adjusted R-squared:  0.6272
## F-statistic: 80.91 on 2 and 93 DF,  p-value: < 2.2e-16
```

However we find the predictor BILL is still not significant, therefore we are going use TYPE*BILL as the predictor and finally both the predictor and intercept are significant.

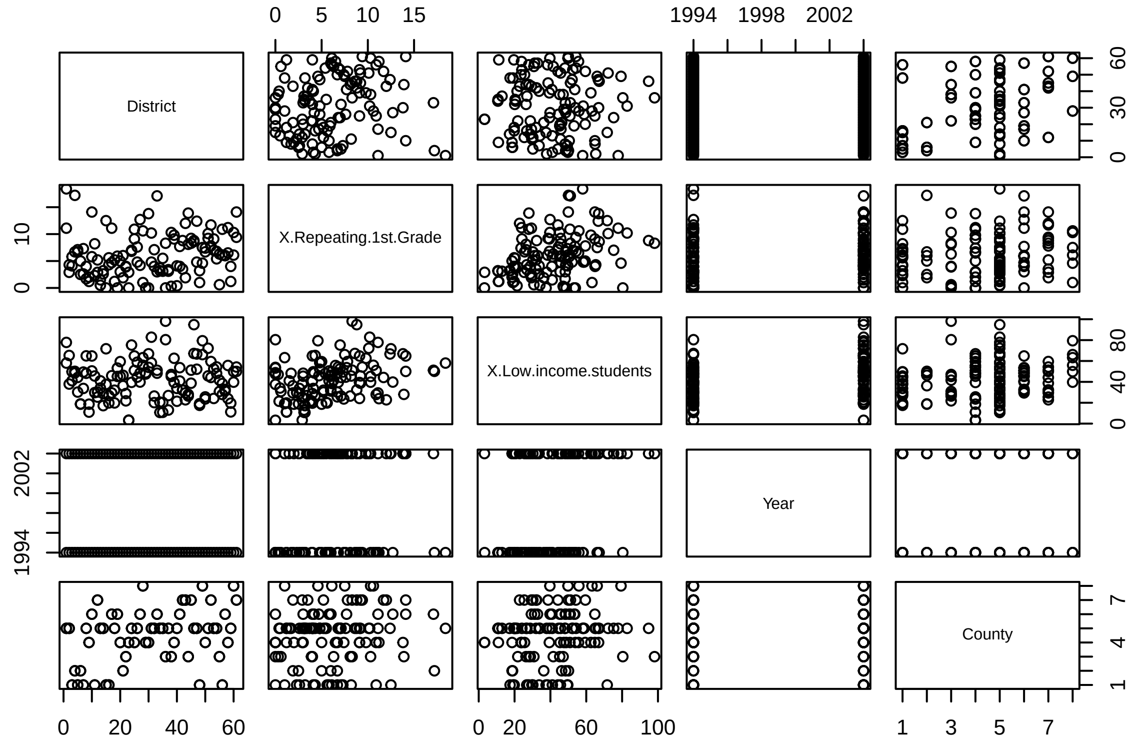
```
overdue_model3 <- lm(LATE ~  TYPE*BILL)
summary(overdue_model3)
```

```
##
## Call:
## lm(formula = LATE ~ TYPE * BILL)
##
## 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) -97.338937   2.679936  -36.32   <2e-16 ***
## TYPE         99.548561   1.694940   58.73   <2e-16 ***
## BILL          0.522327   0.014054   37.17   <2e-16 ***
## TYPE:BILL    -0.356644   0.008888  -40.12   <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
```

Problem 2

(a) True

```
HoustonChron <- read.csv("HoustonChronicle.csv")
attach(HoustonChron)
plot(HoustonChron)
```



```r
cor(HoustonChron$X.Low.income.students, HoustonChron$X.Repeating.1st.Grade)
```

```
## [1] 0.3535689
```

(b) False. Since the P-value is 0.1021, we fail to reject the null hypothesis at 5% level of significance, therefore there is no enough evidence to show a increase in the percentage of students repeating first grade between 1994–1995 and 2004–2005.

```r
HoustonChron$Year <- factor(HoustonChron$Year)
t.test(X.Repeating.1st.Grade~Year, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  X.Repeating.1st.Grade by Year
## t = -1.6474, df = 120, p-value = 0.1021
## alternative hypothesis: true difference in means between group 1994 and group 2004 is not equal to 0
## 95 percent confidence interval:
##  -2.653031  0.243195
## sample estimates:
## mean in group 1994 mean in group 2004
##           5.473770           6.678689
```

(c) Since the P-value is 0.07905, we fail to reject the null hypothesis at 5% level of significance, therefore there is no enough evidence to show there association between percentage of low income student and percentage of students repeating first grade in 1994.

```
x1994 <- HoustonChron$X.Repeating.1st.Grade[HoustonChron$Year == "1994"]
y1994 <- HoustonChron$X.Low.income.students[HoustonChron$Year == "1994"]
x2004 <- HoustonChron$X.Repeating.1st.Grade[HoustonChron$Year == "2004"]
y2004 <- HoustonChron$X.Low.income.students[HoustonChron$Year == "2004"]
cor.test(x1994, y1994)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x1994 and y1994
## t = 1.7872, df = 59, p-value = 0.07905
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02673009  0.45260798
## sample estimates:
##      cor
## 0.226616
```

Since the P-value is 0.0005413, we can to reject the null hypothesis at 5% level of significance, therefore there is enough evidence to show there association between percentage of low income student and percentage of students repeating first grade in 2004.

```
cor.test(x2004, y2004)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x2004 and y2004
## t = 3.6593, df = 59, p-value = 0.0005413
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1999196 0.6152717
## sample estimates:
##      cor
## 0.430088
```

Probelm 3

(a) P-value of EndofHarvest*Rain is 0.0120, therefore we have enough evidence to reject the null hypothesis and conclude that there are enough evidence to show that interaction term in model is significant.

```
latour <- read.table("latour.txt", header = TRUE)
attach(latour)
latour_model <- lm(Quality ~ EndofHarvest + Rain + EndofHarvest*Rain)
summary(latour_model)
```

```
##
```

```
## Call:
## lm(formula = Quality ~ EndofHarvest + Rain + EndofHarvest * Rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6833 -0.5703  0.1265  0.4385  1.6354
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.16122    0.68917   7.489 3.95e-09 ***
## EndofHarvest      -0.03145    0.01760  -1.787   0.0816 .
## Rain               1.78670    1.31740   1.356   0.1826
## EndofHarvest:Rain -0.08314    0.03160  -2.631   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7578 on 40 degrees of freedom
## Multiple R-squared:  0.6848, Adjusted R-squared:  0.6612
## F-statistic: 28.97 on 3 and 40 DF,  p-value: 4.017e-10
```

(b)

(i) y = 5.16122 - 0.03145EndofHarvest + 1.78670Rain - 0.08314EndofHarvest:Rain, when y = -1, rain = 0

```
latour
```

```
##    Vintage Quality EndofHarvest Rain
## 1     1961     5.0           28    0
## 2     1962     4.0           50    0
## 3     1963     1.0           53    1
## 4     1964     3.0           38    0
## 5     1965     1.0           46    1
## 6     1966     4.0           40    0
## 7     1967     3.0           35    1
## 8     1968     2.0           38    1
## 9     1969     2.0           45    1
## 10    1970     4.0           47    0
## 11    1971     3.0           45    1
## 12    1972     1.0           54    1
## 13    1973     2.0           39    1
## 14    1974     1.0           45    1
## 15    1975     4.0           40    1
## 16    1976     3.0           32    0
## 17    1977     2.0           47    0
## 18    1978     4.0           50    0
## 19    1979     3.0           48    0
## 20    1980     1.0           54    1
## 21    1981     3.0           39    1
## 22    1982     5.0           30    0
## 23    1983     3.0           41    0
## 24    1984     1.0           44    1
## 25    1985     4.0           41    0
## 26    1986     4.0             46     0
## 27    1987     1.0             47     1
## 28    1988     4.0             40     0
## 29    1989     4.0             21     0
## 30    1990     5.0             32     0
## 31    1991     3.0             40     1
## 32    1992     1.0             39     1
## 33    1993     3.0             36     1
## 34    1994     3.5             29     1
## 35    1995     4.0             27     0
## 36    1996     5.0             32     0
## 37    1997     4.0             25     0
## 38    1998     3.5             35     1
## 39    1999     3.5             30     0
## 40    2000     5.0             41     0
## 41    2001     3.5             43     0
## 42    2002     4.0             47     0
## 43    2003     5.0             30     0
## 44    2004     4.0             49     0
```

```
(-1-5.16122)/(-0.03145)
```

```
## [1] 195.9052
```

(ii) when y = -1, rain = 1

```
(-1-5.16122-1.78670)/-(0.03145+0.08314)
```

```
## [1] 69.35963
```