Assignment 3

Lien Dao

4/10/2022

4.2.1

```
library(datasets)
library(readr)
url1 <- "https://gattonweb.uky.edu/sheather/book/docs/datasets/ProfessorSalaries.txt"
salary <- read.table(url1, header = TRUE)
head(salary)
```

```
##   Experience SampleSize ThirdQuartile
## 1          0         17        101300
## 2          2         33        111303
## 3          4         19         98000
## 4          6         25        124000
## 5          8         18        128475
## 6         12         60        117410
```

Simple linear regression model

```
salary_lm <- lm(ThirdQuartile ~ Experience, data = salary)
summary(salary_lm)
```

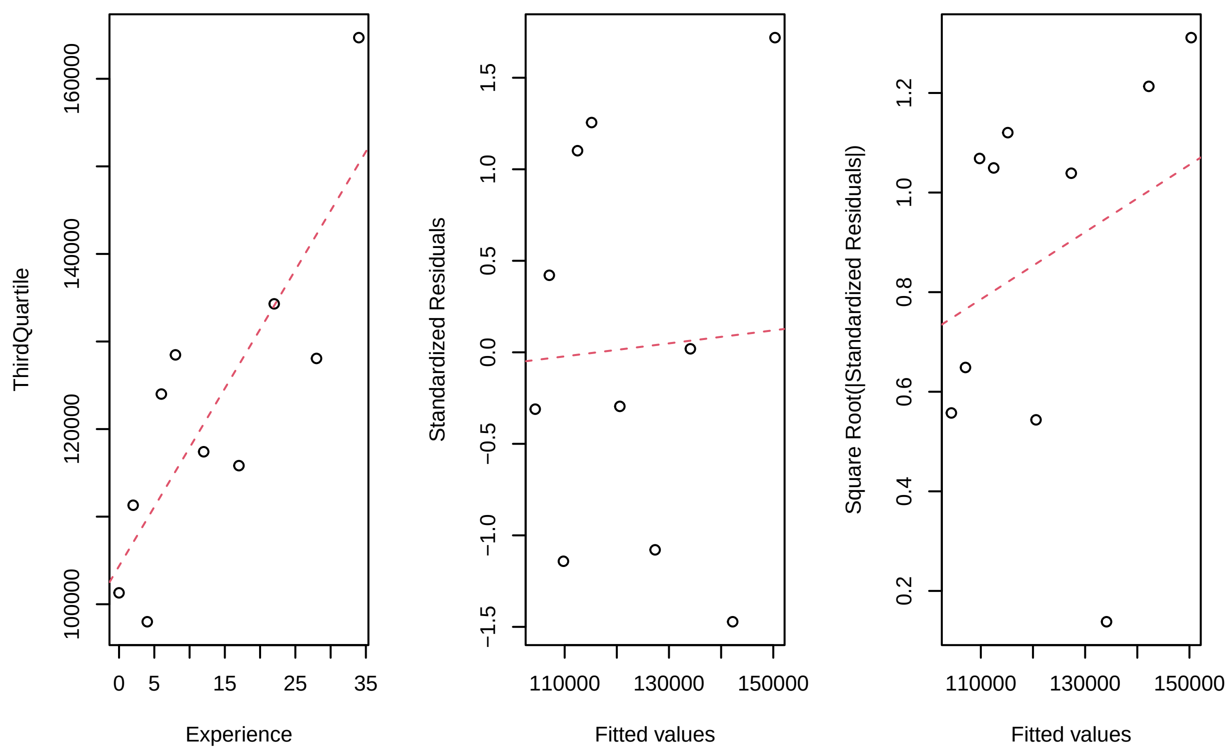
```
##
## Call:
## lm(formula = ThirdQuartile ~ Experience, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14150  -9430  -1428   9712  14370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104352.9     5619.4  18.570 7.29e-08 ***
## Experience    1352.3      325.7   4.152   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11320 on 8 degrees of freedom
## Multiple R-squared:  0.683,  Adjusted R-squared:  0.6434
## F-statistic: 17.24 on 1 and 8 DF,  p-value: 0.0032
```

```
anova(salary_lm)
```

```
## Analysis of Variance Table
```

```
##
## Response: ThirdQuartile
##             Df     Sum Sq    Mean Sq F value Pr(>F)
## Experience  1 2209121299 2209121299  17.239 0.0032 **
## Residuals   8 1025153452  128144182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,3))
plot(salary$Experience, salary$ThirdQuartile, xlab = "Experience",
                                              ylab = "ThirdQuartile")
abline(salary_lm, lty=2,col=2)
fit_val <- fitted(salary_lm)
StanRes1 <- rstandard(salary_lm)
absrtsr1 <- sqrt(abs(StanRes1))
plot(fit_val, StanRes1, ylab="Standardized Residuals", xlab = "Fitted values")
abline(lsfit(fit_val, StanRes1), lty=2, col=2)
plot(fit_val, absrtsr1, ylab="Square Root(|Standardized Residuals|)",
                                  xlab = "Fitted values")
abline(lsfit(fit_val, absrtsr1), lty=2, col=2)
```



When we examine the scatter plot between third quartile salary and experience, we notice that the observations don't follow the line of best fit and one of them is too far from both the line and other points. The distribution of random errors from residual plots shows that the condition for constant variance does not meet since it has a slight funnel shape.

Weighted least squares

```
wt <- 1 / lm(abs(salary_lm$residuals) ~ salary_lm$fitted.values)$fitted.values^2

salary_wls <- lm(ThirdQuartile ~ Experience, data = salary,  weights = wt)
summary(salary_wls)
```

```
##
## Call:
## lm(formula = ThirdQuartile ~ Experience, data = salary, weights = wt)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6750 -1.0474 -0.1576  1.0899  1.6623
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104899.2     4720.8  22.221 1.78e-08 ***
## Experience    1308.9      351.3   3.725  0.00583 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.279 on 8 degrees of freedom
## Multiple R-squared:  0.6343, Adjusted R-squared:  0.5886
## F-statistic: 13.88 on 1 and 8 DF,  p-value: 0.005826
```

```
anova(salary_wls)
```

```
## Analysis of Variance Table
##
## Response: ThirdQuartile
##            Df Sum Sq Mean Sq F value   Pr(>F)
## Experience  1 22.713 22.7135  13.879 0.005826 **
## Residuals   8 13.092  1.6366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Function to calculate weighted parameters

```
wls_stats <- function(n,x,y,wts){
  wmean_x <- sum(wts*x)/sum(wts)
  wmean_y <- sum(wts*y)/sum(wts)

  beta_1_hat = sum(wts*(x-wmean_x)*(y-wmean_y))/sum(wts*(x-wmean_x)^2)
  beta_0_hat = wmean_y-beta_1_hat*wmean_x
  result <- data.frame(n, wmean_x, wmean_y,
                       beta_0_hat, beta_1_hat)
}
wls_result <- wls_stats(10, salary$Experience, salary$ThirdQuartile,wt)
wls_result
```

```
##    n  wmean_x  wmean_y beta_0_hat beta_1_hat
## 1 10 9.488784 117318.9   104899.2   1308.877
```

Predicted y using weighted statistics

```
pred_y <- wls_result[1,4] + wls_result[1,5]*salary$Experience
```

Function to calculate ANOVA statistics using weighted statistics

```
ANOVA <- function(n,x,y,wm_y,y_pred, wt){
  df_reg <- 1
  df_res <- n-2
  df_total <- n-1

  SS_reg = sum(wt*(y_pred - wm_y)^2)
  SSE = sum(wt*(y - y_pred)^2)
  SST = sum(wt*(y - wm_y)^2)

  MSR = SS_reg/df_reg
  MSE = SSE/df_res

  F_stat = MSR/MSE

  p_value <- pf(F_stat, df_reg, df_res, lower.tail = FALSE)
  result <- data.frame(df_reg, df_res, df_total,
                       SS_reg, SSE, SST,
                       MSR, MSE, NA,
                       F_stat, NA, NA,
                       p_value, NA, NA)
}
anova_values <- ANOVA(10,salary$Experience,salary$ThirdQuartile,
                      wls_result[1,3],pred_y,wt)
anova_values
```

```
##   df_reg df_res df_total   SS_reg      SSE      SST      MSR      MSE NA.
## 1      1      8        9 22.71352 13.09253 35.80605 22.71352 1.636566  NA
##     F_stat NA..1 NA..2     p_value NA..3 NA..4
## 1 13.87877    NA    NA 0.005825887    NA    NA
```

The values are the same as using anova().

Estimated third quartile salary of full professors with 6 years of experience

```
sixy_salary <- wls_result[1,4] + wls_result[1,5]*6
```

5.4.2

```
library(readr)
HoustonChronicle <- read_csv("C:/Users/Sen/Downloads/HoustonChronicle.csv")
head(HoustonChronicle)
```

```
## # A tibble: 6 x 5
##   District   `%Repeating 1st Grade` `%Low income students`  Year County
##   <chr>                       <dbl>                  <dbl> <dbl> <chr>
## 1 Alvin                         4.1                   49.7  2004 Brazoria
## 2 Alvin                         5.8                   41.1  1994 Brazoria
## 3 Angleton                      7.1                   44.2  2004 Brazoria
## 4 Angleton                      6.7                   30.2  1994 Brazoria
## 5 Brazosport                    7.3                   49.4  2004 Brazoria
## 6 Brazosport                    2.6                   33.7  1994 Brazoria
```

```
library(tidyverse)
df <- rename(HoustonChronicle, repeat_pct = "%Repeating 1st Grade",
                               low_income = "%Low income students")
```

a)

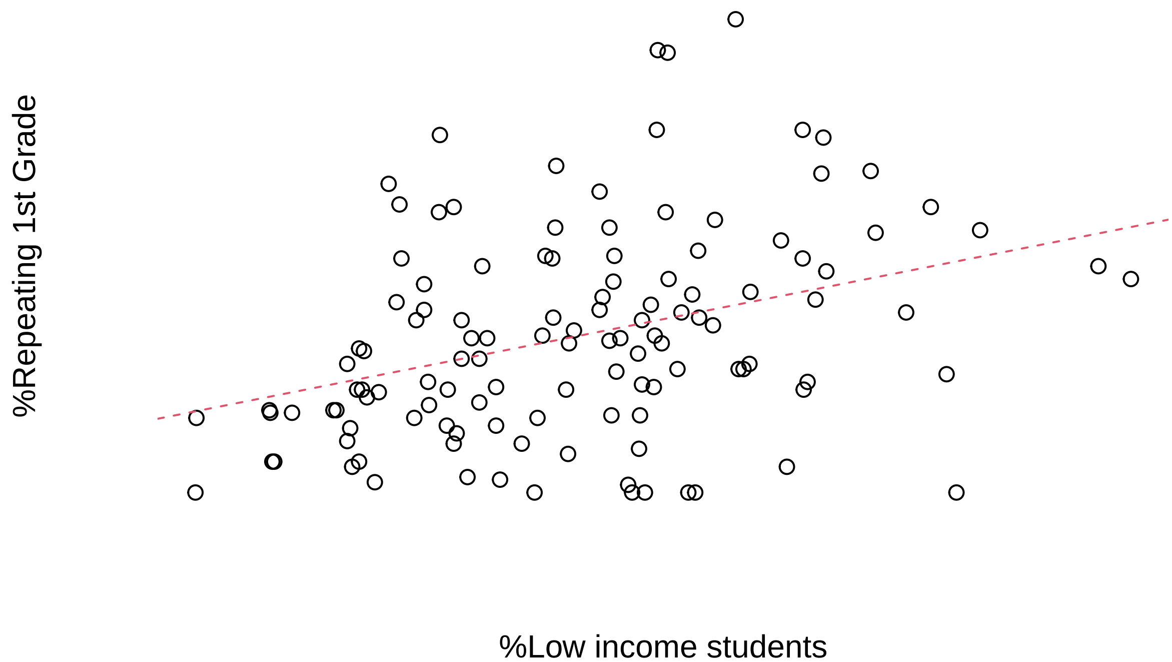
```
m1 <- lm(repeat_pct ~ low_income, data = df)
summary(m1)
```

```
##
## Call:
## lm(formula = repeat_pct ~ low_income, data = df)
##
## 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 ***
## low_income   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
```

```
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: repeat_pct
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## low_income   1  250.29 250.292  17.145 6.472e-05 ***
## Residuals  120 1751.87  14.599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(df$low_income,df$repeat_pct,xlab="%Low income students", ylab="%Repeating 1st Grade", axes=FALSE)
abline(m1, lty=2,col=2)
```

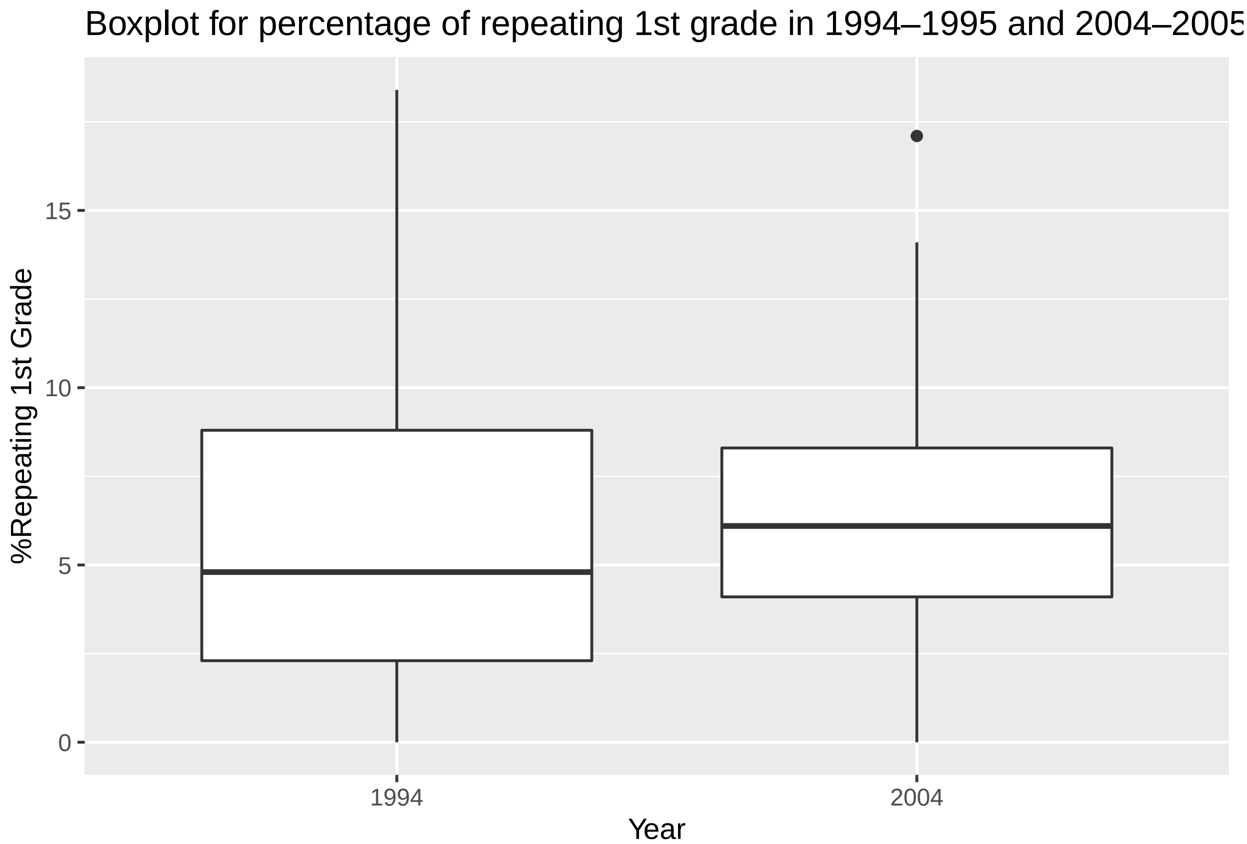


Because the p-value for percentage of low income students is smaller than 0.05 and its coefficient is positive, there is enough evidence to conclude that an increase in the percentage of low income students is associated with an increase in the percentage of students repeating first grade.

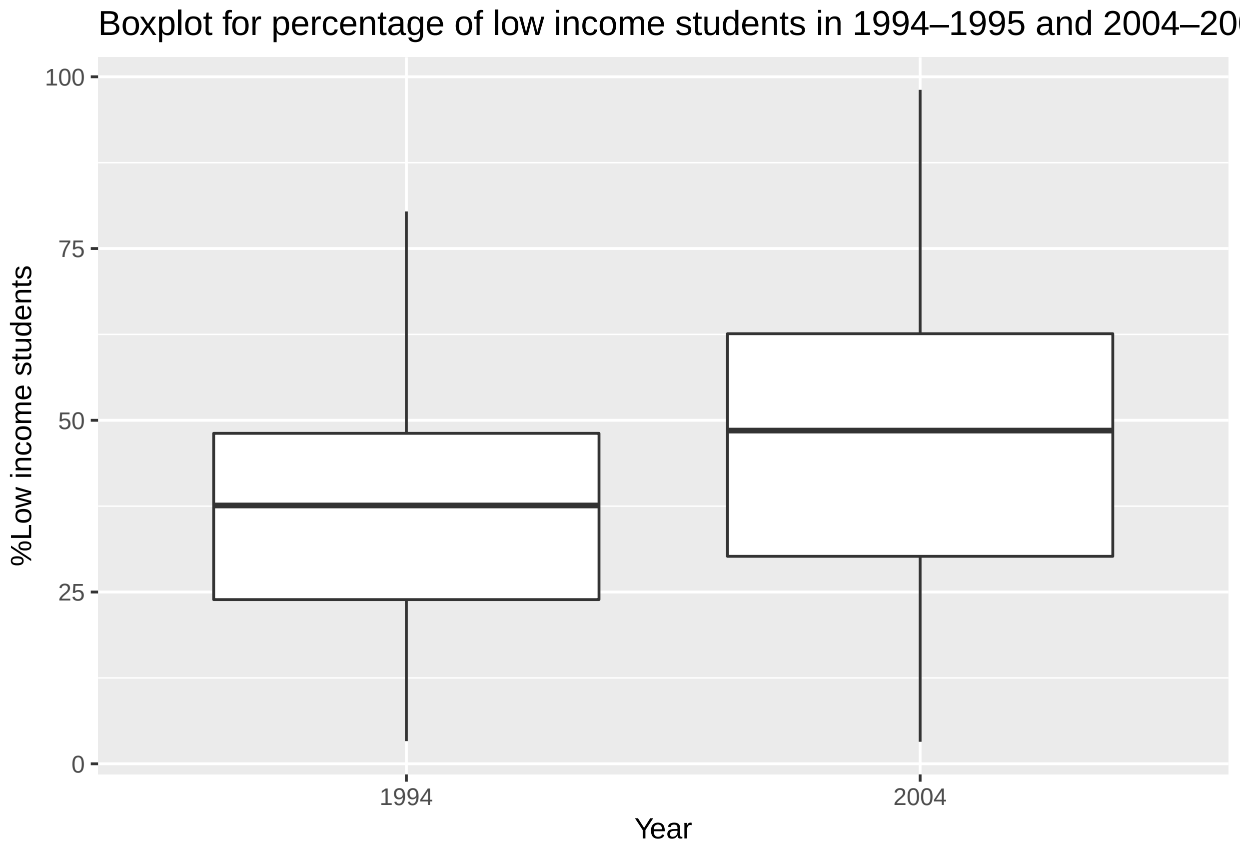
b)

```r
new_year <- as.factor(df$Year)

library(ggplot2)
ggplot(aes(y = df$repeat_pct, x = new_year), data = df) + geom_boxplot() +
       xlab("Year") + ylab("%Repeating 1st Grade") +
       ggtitle("Boxplot for percentage of repeating 1st grade in 1994-1995 and 2004-2005")
```




```
ggplot(aes(y = df$low_income, x = new_year), data = df) + geom_boxplot() +
       xlab("Year") + ylab("%Low income students") +
       ggtitle("Boxplot for percentage of low income students in 1994-1995 and 2004-2005")
```




```
levels(as.factor(new_year))
```

```
## [1] "1994" "2004"
```

```
m2 <- lm(repeat_pct ~ low_income + new_year, data = df)
summary(m2)
```

```
##
## Call:
## lm(formula = repeat_pct ~ low_income + new_year, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6768 -2.5451 -0.4769  1.6624 11.3469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.84900    0.84995   3.352 0.001076 **
## low_income   0.07248    0.01917   3.782 0.000245 ***
## new_year2004 0.38311    0.72716   0.527 0.599274
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.832 on 119 degrees of freedom
## Multiple R-squared:  0.127,  Adjusted R-squared:  0.1124
## F-statistic: 8.659 on 2 and 119 DF,  p-value: 0.0003083
```

```r
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: repeat_pct
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## low_income   1  250.29 250.292 17.0414 6.819e-05 ***
## new_year     1    4.08   4.077  0.2776    0.5993
## Residuals  119 1747.79  14.687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The boxplots of percentage of students repeating first grade and percentage of low income students from 1994 and 2004 both show that the overall percentages in 2004 are higher. Because we choose the period to be a dummy variable, when the students repeated 1st grade in 1994-1995, the model is %Repreat = 2.849 + 0.07248x; but when the students repeated 1st grade in 2004-2005, the model is %Repreat = 3.232 + 0.07248x, which shows that there has been an increase in the percentage of students repeating first grade between 1994–1995 and 2004–2005.

c)

```r
m3 <- lm(repeat_pct ~ low_income + new_year + low_income*new_year, data = df)
summary(m3)
```

```
##
## Call:
## lm(formula = repeat_pct ~ low_income + new_year + low_income *
##     new_year, data = df)
##
## 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)              3.27194    1.22347   2.674  0.00855 **
## low_income               0.06080    0.03093   1.966  0.05167 .
## new_year2004            -0.38956    1.76109  -0.221  0.82532
## low_income:new_year2004  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
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: repeat_pct
##                      Df  Sum Sq Mean Sq F value    Pr(>F)
## low_income            1  250.29 250.292 16.9314 7.208e-05 ***
## new_year              1    4.08   4.077  0.2758    0.6005
## low_income:new_year   1    3.44   3.435  0.2324    0.6307
```

```
## Residuals           118 1744.36  14.783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m1,m3)
```

```
## Analysis of Variance Table
##
## Model 1: repeat_pct ~ low_income
## Model 2: repeat_pct ~ low_income + new_year + low_income * new_year
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    120 1751.9
## 2    118 1744.4  2     7.512 0.2541 0.7761
```

At the 0.05 significant level, there is not enough evidence to support the hypothesis that there is an association between percentage of low income students and the years. Also, given the p-value equals 0.7761 when we compare the reduced model with the model containing the interaction, there is no evidence to support the alternative hypothesis. This means that we will only adopt the reduced model.

5.4.3

```
url2 <- "https://gattonweb.uky.edu/sheather/book/docs/datasets/Latour.txt"
harvest <- read.table(url2, header = TRUE)
head(harvest)
```

```
##   Vintage Quality EndofHarvest Rain
## 1    1961       5           28    0
## 2    1962       4           50    0
## 3    1963       1           53    1
## 4    1964       3           38    0
## 5    1965       1           46    1
## 6    1966       4           40    0
```

a)

```
m4 <- lm(Quality ~ EndofHarvest + Rain + EndofHarvest*Rain, data = harvest)
m4_reduced <- lm(Quality ~ EndofHarvest + Rain, data = harvest)
summary(m4)
```

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest + Rain + EndofHarvest * Rain,
##     data = harvest)
##
## 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
```

```
anova(m4_reduced,m4)
```

```
## Analysis of Variance Table
##
## Model 1: Quality ~ EndofHarvest + Rain
## Model 2: Quality ~ EndofHarvest + Rain + EndofHarvest * Rain
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     41 26.945
## 2     40 22.971  1    3.9749 6.9218 0.01203 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value of the interaction variable between the days and rain is less than 0.05, the coefficient of the interaction term is statistically significant, which shows that the rate of change in quality rating depends on whether there has been any unwanted rain at vintage.

b) Estimate the number of days of delay to the end of harvest it takes to decrease the quality rating by 1 point

```
coeff <- m4$coefficients
coeff
```

```
##       (Intercept)      EndofHarvest              Rain EndofHarvest:Rain
##        5.16121899       -0.03144552        1.78669768       -0.08313781
```

(i) No unwanted rain at harvest

```
no_rain_day <- ((coeff[1]-1)-coeff[1])/coeff[2]
no_rain_day
```

```
## (Intercept)
##    31.80103
```

(ii) Some unwanted rain at harvest

```
some_rain_day <- ((coeff[1]+coeff[3]-1)-coeff[1]-coeff[3])/(coeff[2]+coeff[4])
some_rain_day
```

```
## (Intercept)
##    8.727273
```

End