STAT 757 Assignment 5 Solutions

DUE 4/01/2018 11:59PM

AG Schissler

2/14/2018

Instructions [20 points]

Modify this file to provide responses to the Ch.5 Exercises in Sheather (2009). You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter5.R. Also address the project milestones indicated below. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format SURNAME-FIRSTNAME-Assignment5.Rmd and SURNAME-FIRSTNAME-Assignment5.pdf.

```
data_dir <- "/Users/alfred/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data"
```

Exercise 5.4.2

```
reading <- read.csv(file.path(data_dir, "HoustonChronicle.csv"),header=TRUE)
str(reading)
```

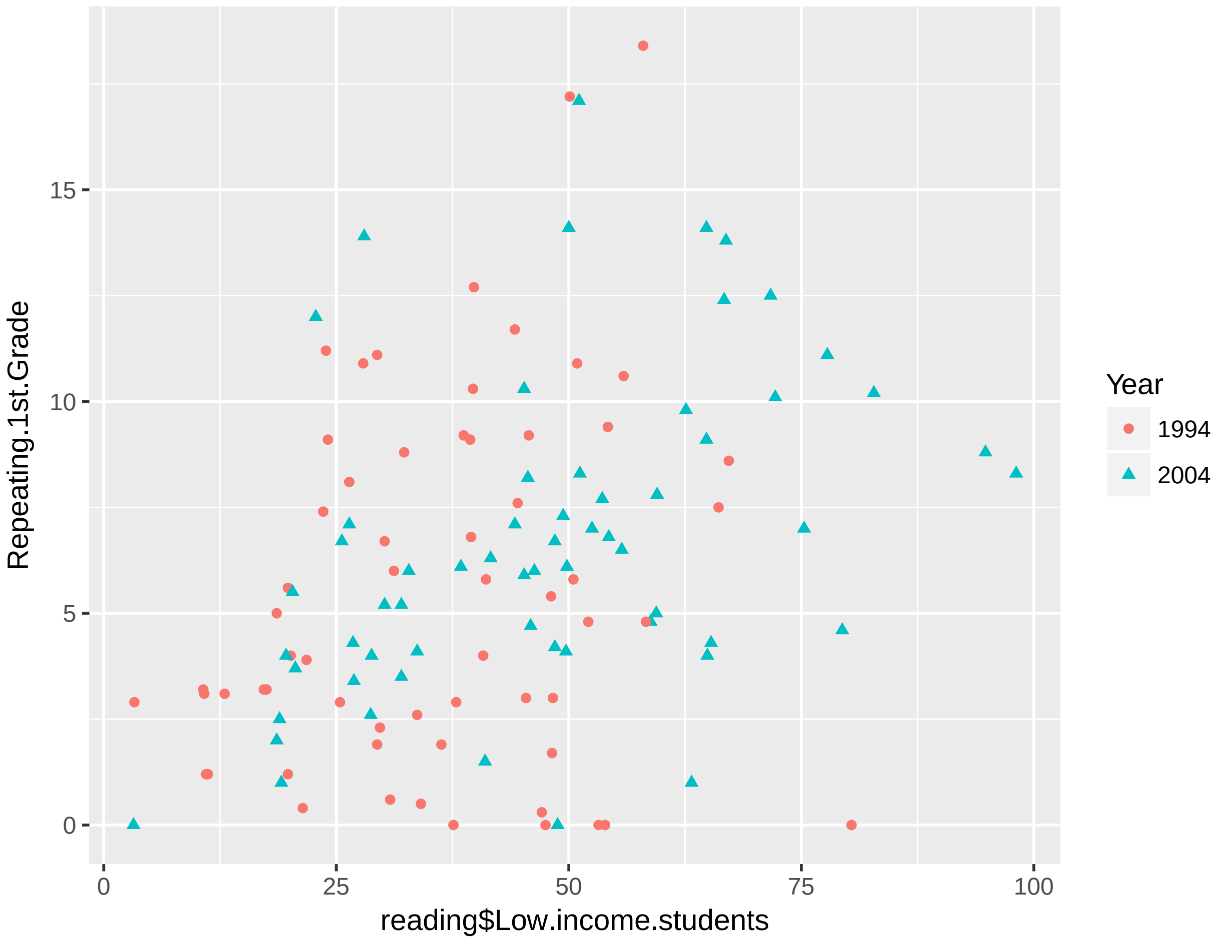
```
## 'data.frame':    122 obs. of  5 variables:
##  $ District             : Factor w/ 61 levels "Aldine","Alief",..: 3 3 5 5 7 7 11 11 15 15 ...
##  $ X.Repeating.1st.Grade: num  4.1 5.8 7.1 6.7 7.3 2.6 8.2 2.3 12.5 0 ...
##  $ X.Low.income.students: num  49.7 41.1 44.2 30.2 49.4 33.7 45.6 29.7 71.7 37.6 ...
##  $ Year                 : int  2004 1994 2004 1994 2004 1994 2004 1994 2004 1994 ...
##  $ County               : Factor w/ 8 levels "Brazoria","Chambers",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(reading)
```

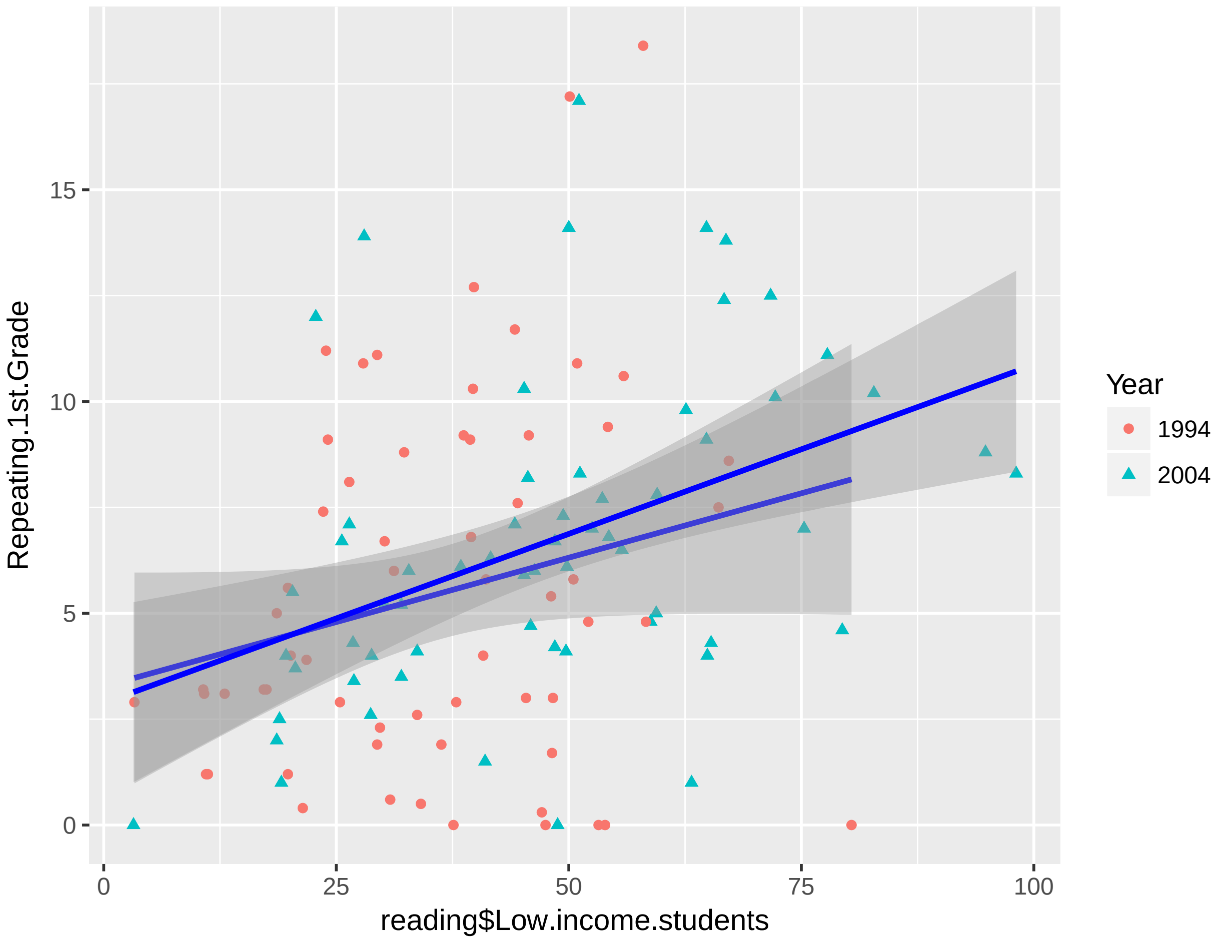
```
##     District X.Repeating.1st.Grade X.Low.income.students Year   County
## 1      Alvin                   4.1                  49.7 2004 Brazoria
## 2      Alvin                   5.8                  41.1 1994 Brazoria
## 3   Angleton                   7.1                  44.2 2004 Brazoria
## 4   Angleton                   6.7                  30.2 1994 Brazoria
## 5 Brazosport                   7.3                  49.4 2004 Brazoria
## 6 Brazosport                   2.6                  33.7 1994 Brazoria
```

```
## clean up names
names(reading) <- gsub("X\\.", "", names(reading))
## "factorize" year
reading$Year <- factor(reading$Year)
```

```
require(ggplot2)
p1 <- ggplot(data = reading, aes(x = reading$Low.income.students, y = Repeating.1st.Grade, shape = Year
p1 + geom_point()
```




```
p1 + geom_point() + stat_smooth(method = "lm", col = "blue")
```



The Year category seems fairly randomly disributed across data points. But this may be something that is hard to see by eye.

5.4.2 Part a

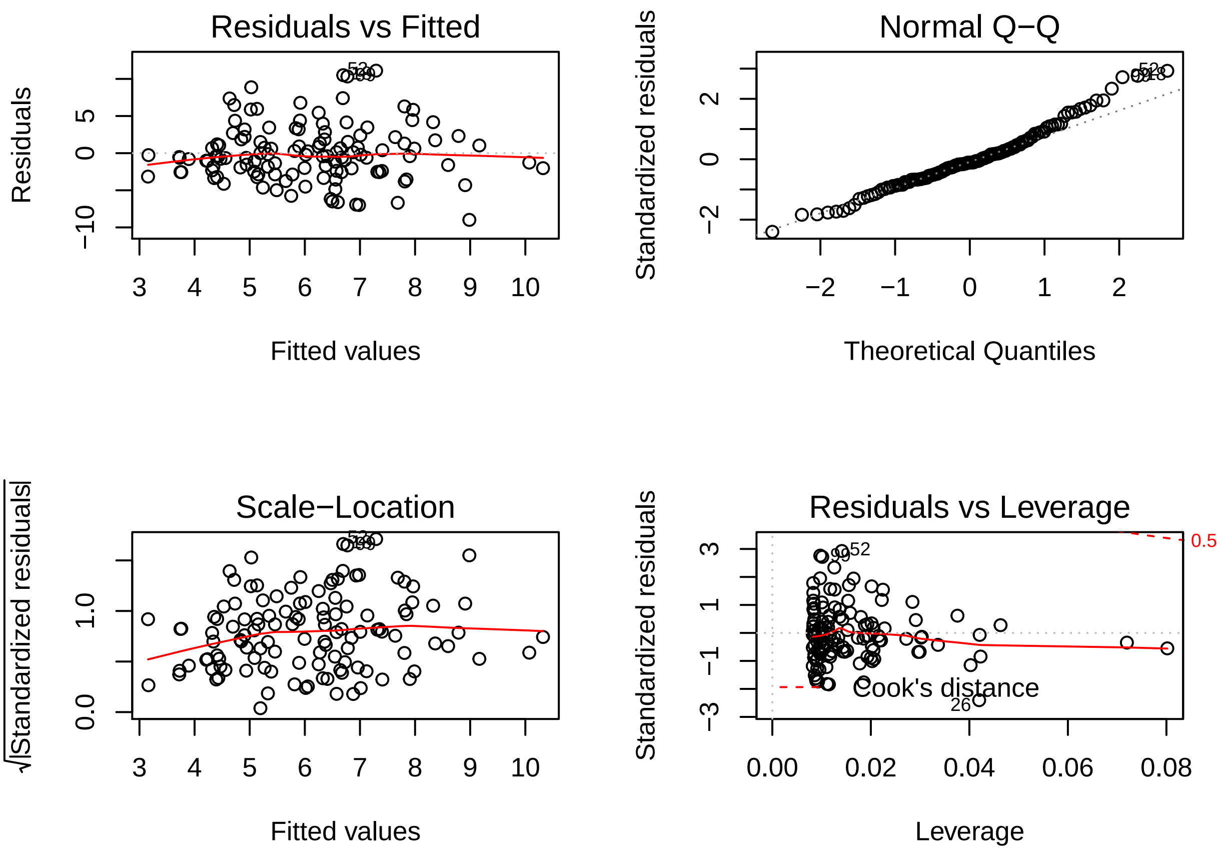
This is a the analysis of covariance (ANCOVA) scenario with co-incident lines.

```
fit1 <- lm(Repeating.1st.Grade ~ Low.income.students, data = reading)
summary(fit1)
```

```
##
## Call:
## lm(formula = Repeating.1st.Grade ~ Low.income.students, data = reading)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.985 -2.507 -0.418  1.850 11.107
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.9142     0.8384    3.48  0.00071
## Low.income.students   0.0755     0.0182    4.14  6.5e-05
##
## Residual standard error: 3.82 on 120 degrees of freedom
## Multiple R-squared:  0.125,  Adjusted R-squared:  0.118
```

```
## F-statistic: 17.1 on 1 and 120 DF,  p-value: 6.47e-05
```

```r
par(mfrow=c(2,2))
plot(fit1)
```



Looks like a decent fit and the slope is significantly greater than 0.

5.4.2 Part b

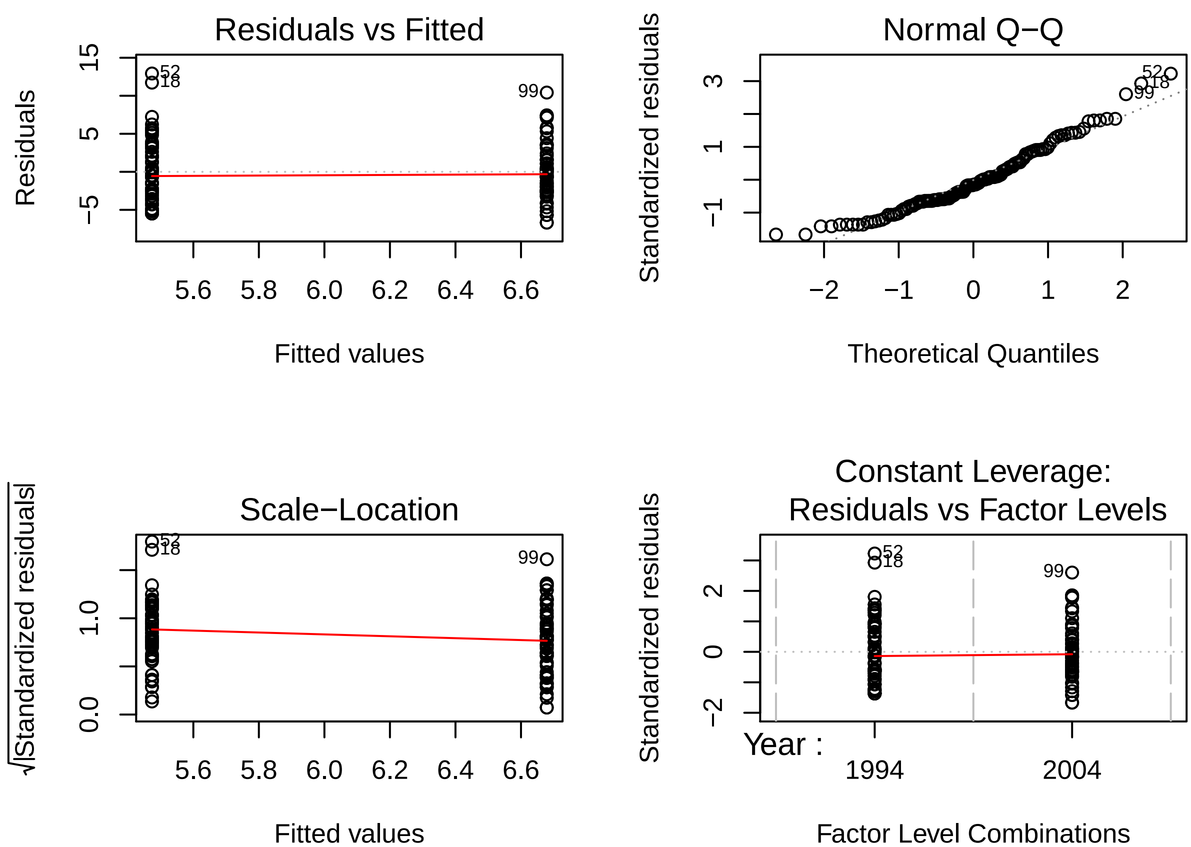
This scenario ignores the low-income percentage.

```r
fit2 <- lm(Repeating.1st.Grade ~ Year, data = reading)
summary(fit2)
```

```
##
## Call:
## lm(formula = Repeating.1st.Grade ~ Year, data = reading)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.679 -2.654 -0.626  2.575 12.926
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.474      0.517   10.58   <2e-16
## Year2004       1.205      0.731    1.65      0.1
##
## Residual standard error: 4.04 on 120 degrees of freedom
```

```
## Multiple R-squared:  0.0221, Adjusted R-squared:  0.014
## F-statistic: 2.71 on 1 and 120 DF,  p-value: 0.102
```

```r
par(mfrow=c(2,2))
plot(fit2)
```



This amounts to a t-test on the different means between the Years. There is not much evidence of an increase when not accounting for low-income status.

5.4.2 Part c

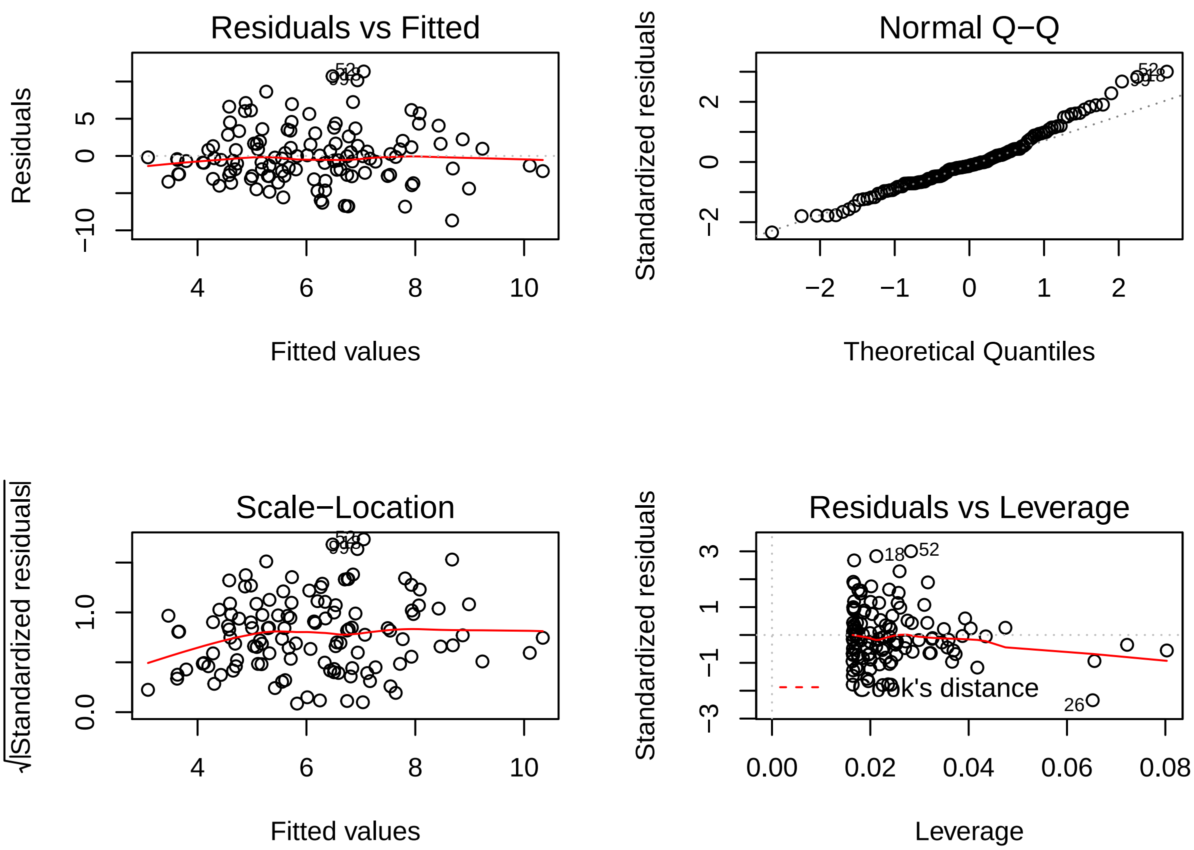
The scenario below assumes an additive effect of Year (parallel lines).

```r
fit3 <- lm(Repeating.1st.Grade ~ Year + Low.income.students, data = reading)
summary(fit3)
```

```
##
## Call:
## lm(formula = Repeating.1st.Grade ~ Year + Low.income.students,
##     data = reading)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.677 -2.545 -0.477  1.662 11.347
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.8490     0.8500    3.35  0.00108
```

```
## Year2004                0.3831     0.7272    0.53  0.59927
## Low.income.students     0.0725     0.0192    3.78  0.00024
##
## Residual standard error: 3.83 on 119 degrees of freedom
## Multiple R-squared:  0.127,  Adjusted R-squared:  0.112
## F-statistic: 8.66 on 2 and 119 DF,  p-value: 0.000308
```

```
par(mfrow=c(2,2))
plot(fit3)
```



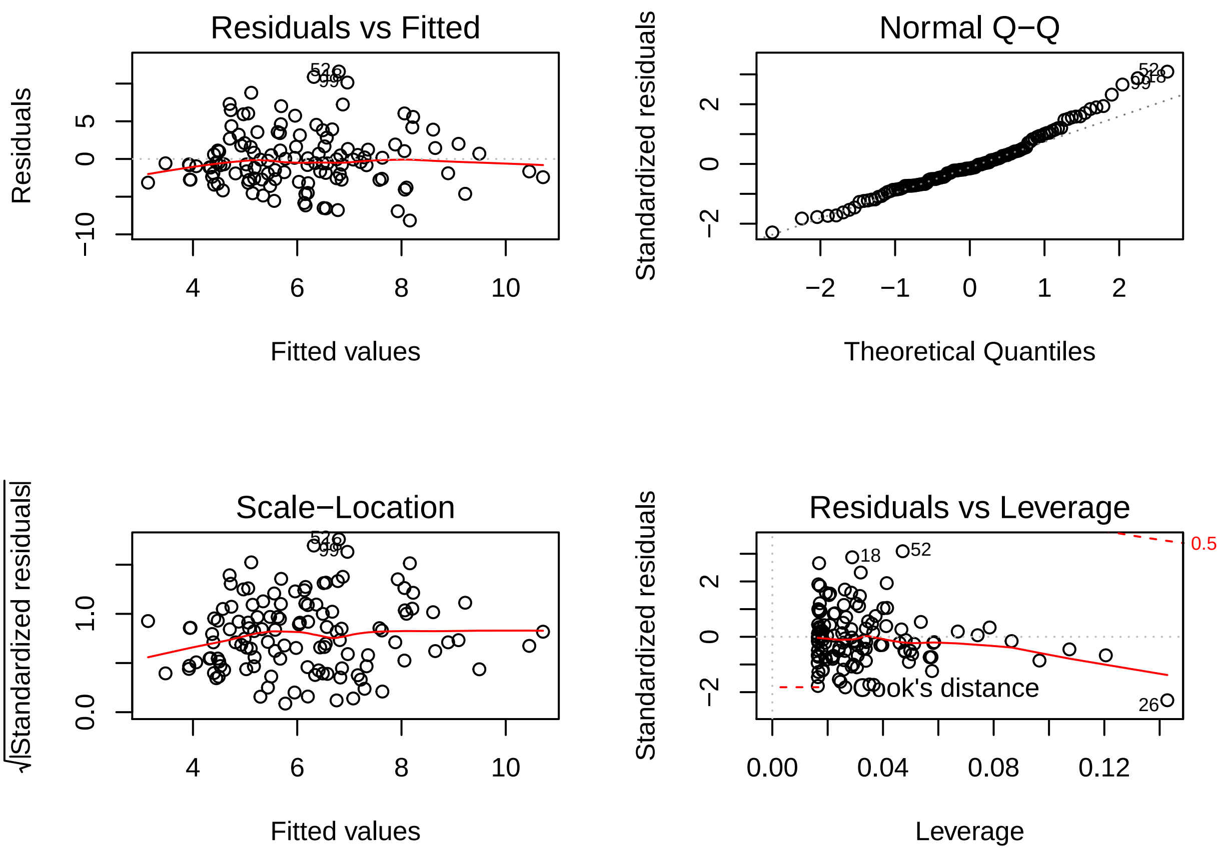
This scenario is the most general ANCOVA.

```
full <- lm(Repeating.1st.Grade ~ Low.income.students + Year + Low.income.students:Year, data = reading)
summary(full)
```

```
##
## Call:
## lm(formula = Repeating.1st.Grade ~ Low.income.students + Year +
##     Low.income.students:Year, data = reading)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.161 -2.612 -0.558  1.750 11.601
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.2719     1.2235    2.67   0.0086
## Low.income.students              0.0608     0.0309    1.97   0.0517
```

```
## Year2004                            -0.3896     1.7611   -0.22   0.8253
## Low.income.students:Year2004   0.0190     0.0395    0.48   0.6307
##
## Residual standard error: 3.84 on 118 degrees of freedom
## Multiple R-squared:  0.129,  Adjusted R-squared:  0.107
## F-statistic: 5.81 on 3 and 118 DF,  p-value: 0.000969
```

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



There is little evidence that the association between the percentage low-income students and percentage reading failure rates differs between the years 1994 and 2004. Let's do an ANOVA to compare models:

```r
## check against only parallel lines model
anova(fit3, full)
```

```
## Analysis of Variance Table
##
## Model 1: Repeating.1st.Grade ~ Year + Low.income.students
## Model 2: Repeating.1st.Grade ~ Low.income.students + Year + Low.income.students:Year
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    119 1748
## 2    118 1744  1      3.44 0.23   0.63
```

```r
## check against only Year model
anova(fit2, full)
```

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

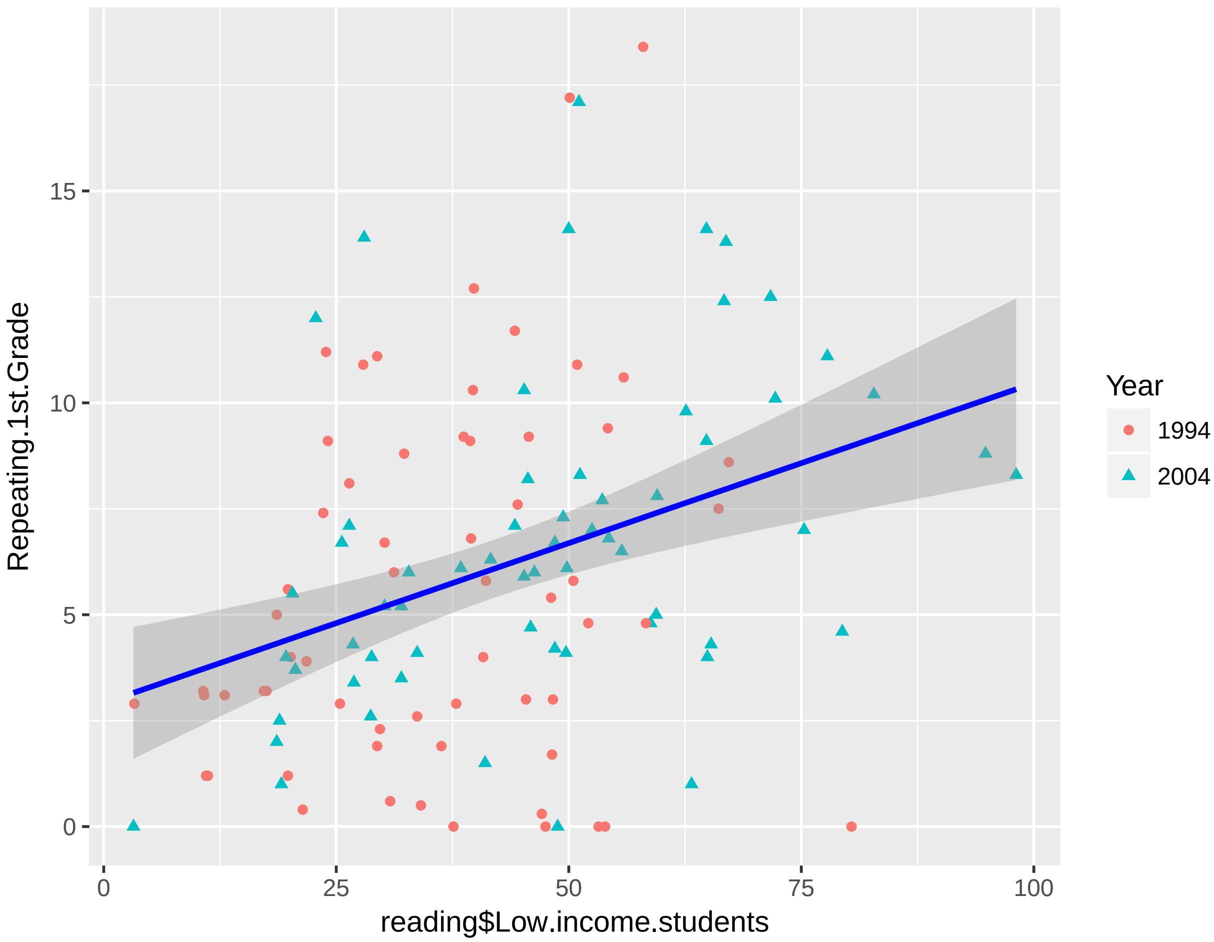
```
## Model 1: Repeating.1st.Grade ~ Year
## Model 2: Repeating.1st.Grade ~ Low.income.students + Year + Low.income.students:Year
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    120 1958
## 2    118 1744  2       214 7.22 0.0011
```

```
## check against only low income model
anova(fit1, full)
```

```
## Analysis of Variance Table
##
## Model 1: Repeating.1st.Grade ~ Low.income.students
## Model 2: Repeating.1st.Grade ~ Low.income.students + Year + Low.income.students:Year
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    120 1752
## 2    118 1744  2      7.51 0.25   0.78
```

Based on this it is reasonable to not consider the Year in the analysis. It does however show including the an interaction with low-income is better than leaving it out entirely (Year-only model). My final analysis would be that the the association does not change with year, but there is a "positive" association between the percentage of low-income students and higher reading test failure rates. The final model is plotting below:

```
require(ggplot2)
p1 <- ggplot(data = reading, aes(x = reading$Low.income.students, y = Repeating.1st.Grade))
p1 + geom_point(aes(shape = Year, color = Year)) + stat_smooth(method = "lm", col = "blue")
```



Exercise 5.4.3

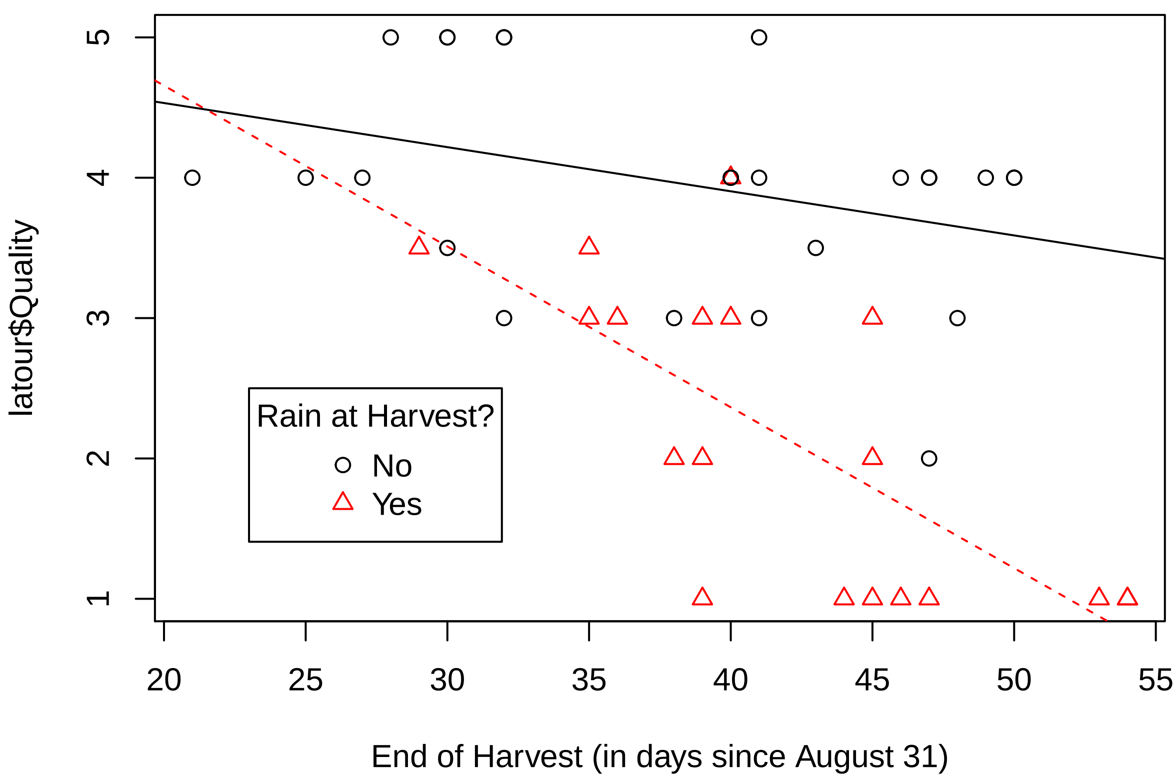
```r
latour <- read.table(file.path(data_dir, "Latour.txt"), header=TRUE)
head(latour)
```

```
##   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
```

```r
str(latour)
```

```
## 'data.frame':    44 obs. of  4 variables:
##  $ Vintage     : int  1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 ...
##  $ Quality     : num  5 4 1 3 1 4 3 2 2 4 ...
##  $ EndofHarvest: int  28 50 53 38 46 40 35 38 45 47 ...
##  $ Rain        : int  0 0 1 0 1 0 1 1 1 0 ...
```

```r
#Figure 5.8 on page 149
y <- latour$Rain
par(mfrow=c(1,1))
plot(latour$EndofHarvest, latour$Quality,pch=y+1,col=y+1,xlab="End of Harvest (in days since August 31)
abline(lsfit(latour$EndofHarvest[y==0],latour$Quality[y==0]),lty=1,col=1)
abline(lsfit(latour$EndofHarvest[y==1],latour$Quality[y==1]),lty=2,col=2)
legend(23, 2.5,legend=c("No","Yes"),pch=1:2,col=1:2,title="Rain at Harvest?")
```



Exercise 5.4.3 Part A

The plot suggests differing slopes and intercepts when considering the rate of change in wine quality and days after August 31 for years with rain at harvest. Note that the y-intercept is not illustrated in the plot. As instructed in the prompt, first let's fit the most general model:

```r
#Regression output on page 148
mfull <- lm(Quality ~ EndofHarvest + Rain + Rain:EndofHarvest, data = latour)
summary(mfull)
```

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest + Rain + Rain:EndofHarvest,
##     data = latour)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.683 -0.570  0.127  0.439  1.635
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.1612     0.6892    7.49  3.9e-09
## EndofHarvest       -0.0314     0.0176   -1.79    0.082
## Rain                1.7867     1.3174    1.36    0.183
## EndofHarvest:Rain  -0.0831     0.0316   -2.63    0.012
##
## Residual standard error: 0.758 on 40 degrees of freedom
## Multiple R-squared:  0.685,  Adjusted R-squared:  0.661
## F-statistic:   29 on 3 and 40 DF,  p-value: 4.02e-10
```

```r
#Regression output on page 149
mreduced <- lm(Quality ~ EndofHarvest + Rain, data = latour)
summary(mreduced)
```

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest + Rain, data = latour)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.456 -0.737  0.143  0.641  1.765
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.1463     0.6190    9.93  1.8e-12
## EndofHarvest  -0.0572     0.0156   -3.66  0.00071
## Rain          -1.6222     0.2548   -6.37  1.3e-07
##
## Residual standard error: 0.811 on 41 degrees of freedom
## Multiple R-squared:  0.63,   Adjusted R-squared:  0.612
## F-statistic:   35 on 2 and 41 DF,  p-value: 1.38e-09
```

```r
anova(mreduced,mfull)
```

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

```
## Model 1: Quality ~ EndofHarvest + Rain
## Model 2: Quality ~ EndofHarvest + Rain + Rain:EndofHarvest
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     41 26.9
## 2     40 23.0  1      3.97 6.92  0.012
```

By inspecting the anova analysis, it is clear that the impact of a rain on the rate of change between the wine quality and later harvest is statistically significant.

Exercise 5.4.3 Part B

To estimate the days delay for one point decrease in wine quality, we follow the description on the top of page 141 in Sheather (2009) and note that:

Let $\Delta Y = Y_2 - Y_1$ and $\Delta x = x_2 - x_1$. And so the question asks for the expected number of days (Δx) with $\Delta Y = -1$.

$$
\begin{aligned}
E(Y|d=0) = \beta_0 + \beta_1 x \Rightarrow \\
E(\Delta Y|d=0) &= E(Y_2 - Y_1|d=0)\beta_0 + \beta_1 x_2 - \beta_0 - \beta_1 x_1 \\
-1 &= \beta_1 \Delta x \\
\frac{-1}{\beta_1} &= \Delta x
\end{aligned}
$$

In the absence of rain, we estimate the days delayed to correspond with a point decrease in quality to be

```r
unname(-1/coefficients(mfull)[2])
```

```
## [1] 31.801
```

Similarly,

$$
\begin{aligned}
E(Y|d=1) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x \Rightarrow \\
E(\Delta Y|d=1) &= (\beta_1 + \beta_3)\Delta x \\
-1 &= (\beta_1 + \beta_3)\Delta x \\
\frac{-1}{\beta_1 + \beta_3} &= \Delta x
\end{aligned}
$$

In the presence of rain, we estimate the days delayed to correspond with a point decrease in quality to be

```r
unname(-1/(coefficients(mfull)[2] + coefficients(mfull)[4]))
```

```
## [1] 8.7273
```

Project milestones [20 points]

1. Perform an exploratory data analysis:

- Numerically summarize the variables.
- Make plots and explore relationships between variables.
- Identify any strange points or anything else that doesn't make sense.

2. Begin to think about how to model the relationships in your data.

References

Sheather, Simon. 2009. *A Modern Approach to Regression with R*. Springer Science & Business Media.