Assignment 5

Biplav Timalsina

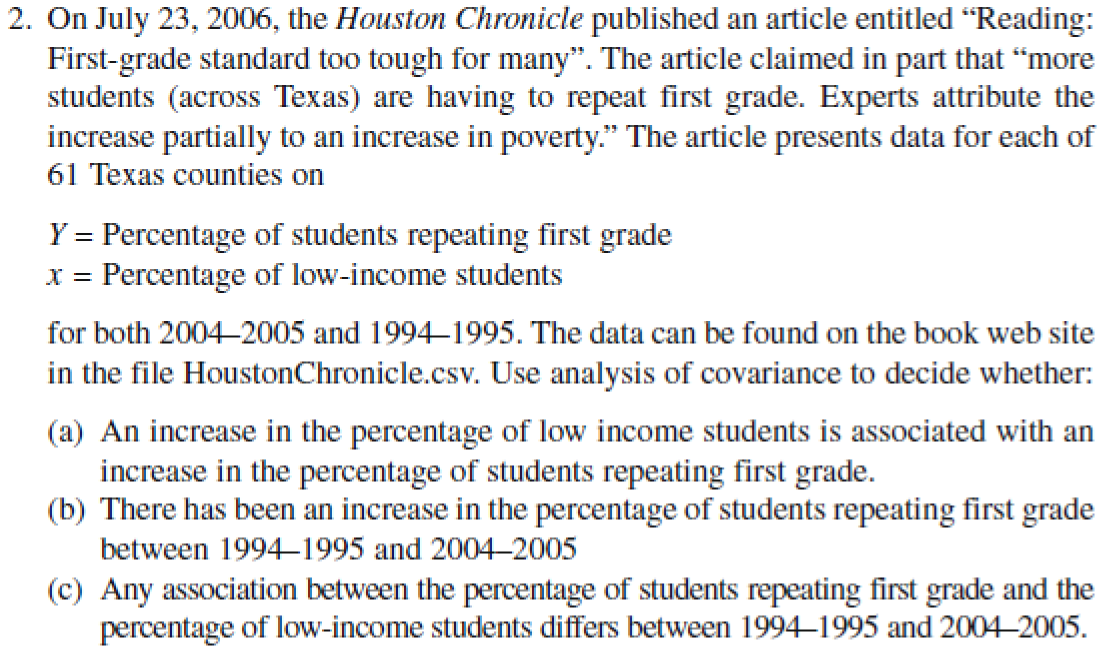
April 01, 2018

STAT 757 Applied Regression Analysis

Instructions [20 points]

Modify this file to provide responses to the Ch.5 Exercises in @sheather2009. 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.

Choose either Exercise 5.4.2 or 5.4.3. Respond to all parts posed. [60 points]



2. On July 23, 2006, the *Houston Chronicle* published an article entitled "Reading: First-grade standard too tough for many". The article claimed in part that "more students (across Texas) are having to repeat first grade. Experts attribute the increase partially to an increase in poverty." The article presents data for each of 61 Texas counties on

   Y = Percentage of students repeating first grade
   x = Percentage of low-income students

   for both 2004–2005 and 1994–1995. The data can be found on the book web site in the file HoustonChronicle.csv. Use analysis of covariance to decide whether:

   (a) An increase in the percentage of low income students is associated with an increase in the percentage of students repeating first grade.
   (b) There has been an increase in the percentage of students repeating first grade between 1994–1995 and 2004–2005
   (c) Any association between the percentage of students repeating first grade and the percentage of low-income students differs between 1994–1995 and 2004–2005.

Here, we want to model the response variable, Y which is percentage of students repeating first grade, bases on a continuous predictor, X, which is percentage of low-income students, and a dummy variable year.

We should start with Unrelated regression lines as:

$$Y = \beta_0 + \beta_1 x + e$$

when year=0

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

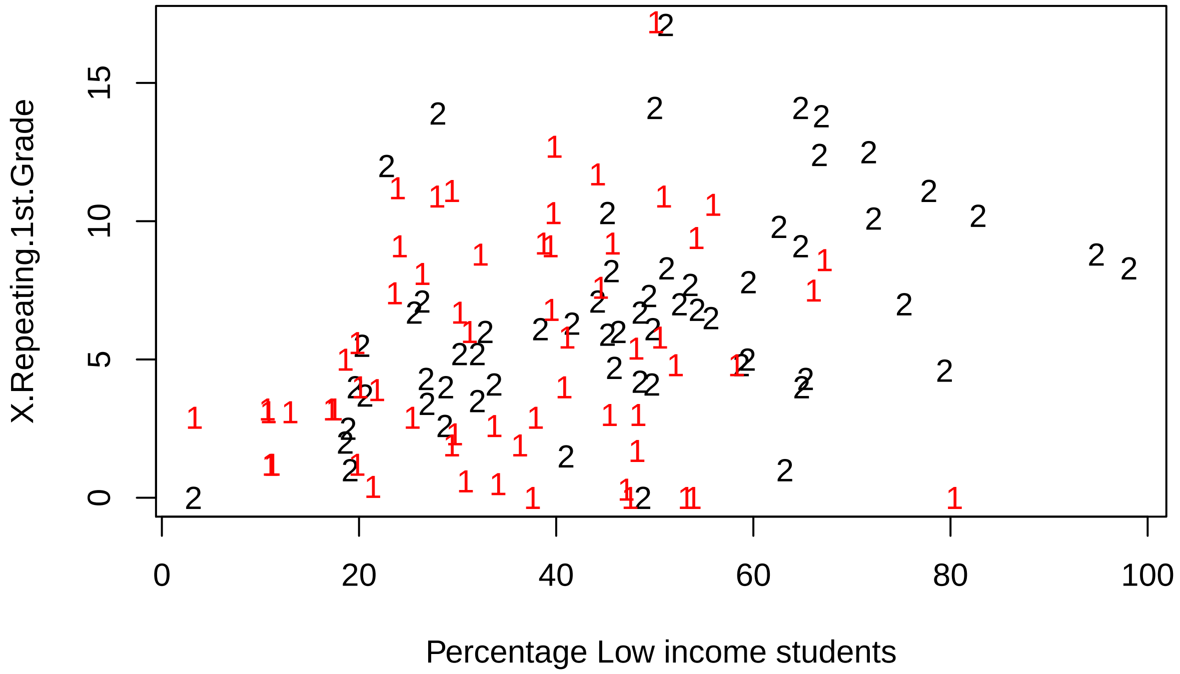
when year = 1

where, Y= percentage of students repeating first grade X= percentage of low income students year= dummy variable (1=2004, 0=1994)

Let us perfrom regression analysis on the data. Let us first plot thed data.

```
houston <- read.csv("F:/unr/4th sem/applied regression analysis/Assignments/HW5/HoustonChronicle.csv",he
attach(houston)

par(mfrow=c(1,1))
plot(X.Low.income.students[Year==2004],X.Repeating.1st.Grade[Year==2004],pch=c("2"),col=c("black"),ylab=
points(X.Low.income.students[Year==1994],X.Repeating.1st.Grade[Year==1994],pch=c("1"),col=c("red"))
```



```
#par(mfrow=c(1,1))
#plot(houston$LowIncome[Year==2004],houston$Repeating[Year==2004], pch=c("2"),col=c("black"),ylab="Repe
#points(houston$LowIncome[Year==1994],houston$Repeating[Year==1994],pch=c("1"),col=c("red"))
```

We do not see much effect of year(dummy variable) on the percentage of repeating students.

The output from R is as follows.

```
#Regression output
mfull<-lm(X.Repeating.1st.Grade~X.Low.income.students + Year + Year: X.Low.income.students)
summary(mfull)
```

```
##
## Call:
## lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students +
##     Year + Year:X.Low.income.students)
##
## 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)                 80.950443 352.012893   0.230    0.819
## X.Low.income.students       -3.734749   7.898023  -0.473    0.637
## Year                        -0.038956   0.176109  -0.221    0.825
## X.Low.income.students:Year   0.001903   0.003949   0.482    0.631
##
```

```
## 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
```

Note that none of the regresison coefficients are highly significant. Thus, we can use our final model as

$$Percentage\_Repeating\_1st\_Grade = 80.95 - 3.73 * Percentage\_Low\_Income$$

when year=1994. and,

$$Percentage\_Repeating\_1st\_Grade = 80.95-0.038+(-3.03+.001)x = 80.912-3.029*Percentage\_Low\_Income$$

when year = 2004.

Now, lets test for the analysis of covariance.

The fit for the reduced model is given below:

```
mreduced <- lm(X.Repeating.1st.Grade~X.Low.income.students)
summary(mreduced)
```

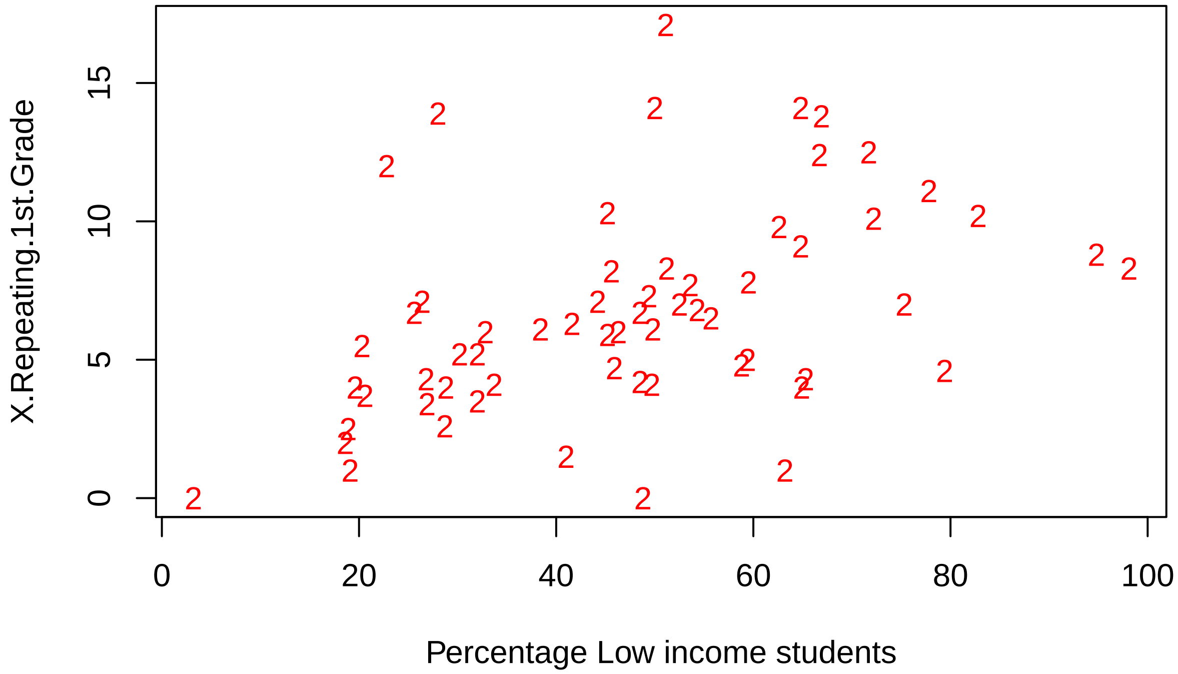
```
##
## Call:
## lm(formula = X.Repeating.1st.Grade ~ X.Low.income.students)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9845 -2.5072 -0.4184  1.8505 11.1067
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.91419    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
```

As expected there is very little evidence that the two models are different.

a) We can see from the negative coefficient of Percentage_Low_Income, that an increase in the percentage of low income students is negatively associated with an increase in percentage of students repeating first grade.

b) To find this let us develop regression models for year 1994 and 2004 seperately.

```
par(mfrow=c(1,1))
plot(X.Low.income.students[Year==2004],X.Repeating.1st.Grade[Year==2004],pch=c("2"),col=c("red"),ylab="
```

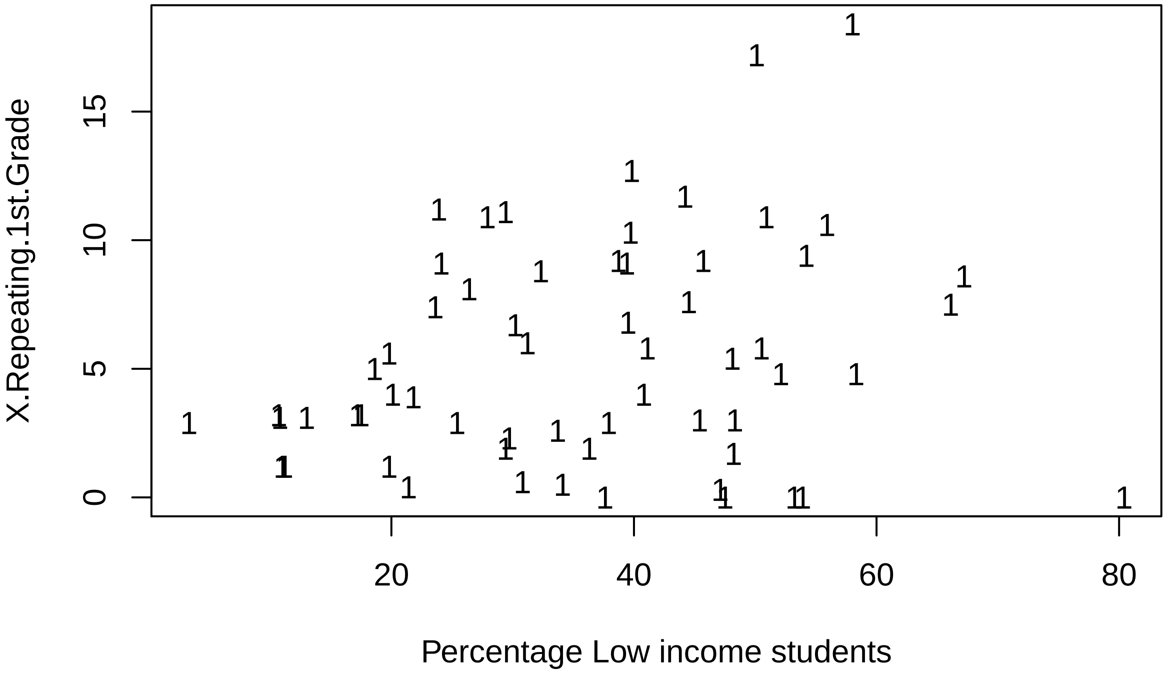


```r
#points(X.Low.income.students[Year==1994],X.Repeating.1st.Grade[Year==1994],pch=c("1"),col=c("red"))

model4<-lm(X.Repeating.1st.Grade[Year==2004]~X.Low.income.students[Year==2004])
summary(model4)
```

```
##
## Call:
## lm(formula = X.Repeating.1st.Grade[Year == 2004] ~ X.Low.income.students[Year ==
##     2004])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9282 -1.9372 -0.4177  1.3298 10.1378
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          2.88238    1.12595   2.560 0.013050
## X.Low.income.students[Year == 2004]  0.07984    0.02182   3.659 0.000541
##
## (Intercept)                         *
## X.Low.income.students[Year == 2004] ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.418 on 59 degrees of freedom
## Multiple R-squared:  0.185,  Adjusted R-squared:  0.1712
## F-statistic: 13.39 on 1 and 59 DF,  p-value: 0.0005413
```

```
par(mfrow=c(1,1))
plot(X.Low.income.students[Year==1994],X.Repeating.1st.Grade[Year==1994],pch=c("1"),col=c("black"),ylab=
```



```
#points(X.Low.income.students[Year==1994],X.Repeating.1st.Grade[Year==1994],pch=c("1"),col=c("red"))

model1<-lm(X.Repeating.1st.Grade[Year==1994]~X.Low.income.students[Year==1994])
summary(model1)
```

```
##
## Call:
## lm(formula = X.Repeating.1st.Grade[Year == 1994] ~ X.Low.income.students[Year ==
##     1994])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1606 -2.7778 -0.6975  3.1493 11.6014
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         3.27194    1.34577   2.431   0.0181 *
## X.Low.income.students[Year == 1994] 0.06080    0.03402   1.787   0.0790 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.229 on 59 degrees of freedom
## Multiple R-squared:  0.05135,    Adjusted R-squared:  0.03528
## F-statistic: 3.194 on 1 and 59 DF,  p-value: 0.07905
```

The two models are as follows: y=2.88 + 0.07 x, for year= 2004 y= 3.27 + 0.06 x , for year=1994

Since the intercepts are almost same and slope is also similar, we cannot say that there has been an increase in the percentage of the students repeating first grade between 1994 and 2004.

c)

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

```
## Analysis of Variance Table
##
## Model 1: X.Repeating.1st.Grade ~ X.Low.income.students
## Model 2: X.Repeating.1st.Grade ~ X.Low.income.students + Year + Year:X.Low.income.students
##   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
```

The anova() command clearly shows that removing the year does not significantly affect the fit of the model (F=0.2531, p=0.776). So, there is no difference between the association between the percentage of students repeating first grade and percentage of low-income students between 1994 and 2004.

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