W271 - Applied Regression and Time Series Analysis - Lab2

Ron Cordell, Subhashini Raghunathan, Lei Yang

March 7, 2016

Question 1: Broken Rulers

You have a ruler of length 1 and you choose a place to break it using a uniform probability distribution. Let random variable X represent the length of the left piece of the ruler. X is distributed uniformly in [0, 1]. You take the left piece of the ruler and once again choose a place to break it using a uniform probability distribution. Let random variable Y be the length of the left piece from the second break.

1). Find the conditional expectation of Y given X, $E(Y|X)$.

Given x, the pdf of y is $f(y \mid x) = \frac{1}{x}$, thus

$$E(Y \mid X) = \int_0^x y f(y \mid x) dy = \int_0^x y \frac{1}{x} dy = \frac{1}{x} \times \frac{y^2}{2}\Big|_0^x = \frac{x}{2}$$

2). Find the unconditional expectation of Y . One way to do this is to apply the law of iterated expectations, which states that $E(Y) = E(E(Y|X))$. The inner expectation is the conditional expectation computed above, which is a function of X. The outer expectation finds the expected value of this function.

Apply the law of iterated expectations:

$$E(Y) = E(E(Y \mid X)) = E\big(\frac{X}{2}\big) = \frac{1}{2}E(X) = \frac{1}{2}\int_0^1 x f(x) dx$$

since X is distributed uniformly in $[0, 1]$, then $f(x) = 1$, thus:

$$E(Y) = \frac{1}{2}\int_0^1 x dx = \frac{1}{4}x^2\Big|_0^1 = \frac{1}{4}$$

3). Write down an expression for the joint probability density function of X and Y , $f_{X,Y}(x, y)$.

The joint probability density function of X and Y:

$$f_{X,Y}(x, y) = f(y \mid x)f(x) = \frac{1}{x}$$

4). Find the conditional probability density function of X given Y , $f_{X|Y}$.

For any given $0 < y < 1$, the value of x must be uniformly distributed in the interval of $(y, 1)$, thus the conditional probability density function of X given Y would be:

$$f_{X|Y} = \frac{1}{1-y}$$

5). Find the expectation of X, given that Y is 1/2, $E(X|Y = 1/2)$

In general

$$E(X \mid Y) = \int_y^1 x f_{X|Y} dx = \int_y^1 x \frac{1}{1-y} dx = \frac{1}{1-y}\frac{x^2}{2}\Big|_y^1 = \frac{\frac{1}{2} - \frac{y^2}{2}}{1-y} = \frac{1+y}{2}$$

therefore $E(X \mid Y = \frac{1}{2}) = \frac{1+\frac{1}{2}}{2} = \frac{3}{4}$

Question 2: Investing

Suppose that you are planning an investment in three different companies. The payoff per unit you invest in each company is represented by a random variable. A represents the payoff per unit invested in the first company, B in the second, and C in the third. A, B, and C are independent of each other. Furthermore, $var(A) = 2var(B) = 3var(C)$. You plan to invest a total of one unit in all three companies. You will invest amount a in the first company, b in the second, and c in the third, where $a, b, c \in [0, 1]$ and $a + b + c = 1$. Find, the values of a, b, and c that minimize the variance of your total payoff.

Solution: The total payoff is $TP = aA + bB + cC$, and the variation is

$$var(TP) = var(aA+bB+cC) = a^2var(A)+b^2var(B)+c^2var(C)+2abCov(A,B)+2bcCov(B,C)+2acCov(A,C)$$

since they are 3 different companies, we assume no correlation in payoff per unit between any two of them, considering $var(A) = 2var(B) = 3var(C)$, we then have:

$$var(TP) = (3a^2 + \frac{3}{2}b^2 + c^2)var(C)$$

to minimize $var(TP)$, we need to minimize the coefficient $3a^2 + \frac{3}{2}b^2 + (1 - a - b)^2$ given $c = 1 - a - b$, and take partial derivatives with respect to a and b, and set them to zero:

$$\begin{cases} 4a + b = 1 \\ 2a + 5b = 2 \end{cases}$$

we then have $a = \frac{1}{6}, b = \frac{1}{3}, c = \frac{1}{2}$, intuitively invest more in company with less variation.

Question 3: Turtles

Next, suppose that the lifespan of a species of turtle follows a uniform dis- tribution over $[0, \theta]$. Here, parameter θ represents the unknown maximum lifespan. You have a random sample of n individuals, and measure the lifespan of each individual i to be y_i.

1). Write down the likelihood function, $l(\theta)$ in terms of $y_1, y_2, ..., y_n$.

With each sample y lifespan uniformly distribute in $[0, \theta]$, probability density function is $f(y_i \mid \theta) = \frac{1}{\theta}$, the likelihood function is:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta) = \frac{1}{\theta^n}$$

2). Based on the previous result, what is the maximum-likelihood estimator for θ?

To maxmize $\mathcal{L}(\theta)$ we need θ as small as possible, and yet it must be no smaller than any y_i, and such θ would be $y_{max} = max(y_i)$

3). Let $\hat{\theta}_{ml}$ be the maximum likelihood estimator above. For the simple case that $n = 1$, what is the expectation of $\hat{\theta}_{ml}$, given θ?

When $n = 1$, the MLE of θ becomes $\hat{\theta_{ml}} = y_1$, thus the expectation given θ is:

$$E\big(\hat{\theta_{ml}} \mid \theta\big) = E(y_1 \mid \theta) = \int_0^\theta f(y \mid \theta) y dy = \int_0^\theta \frac{1}{\theta} y dy = \frac{1}{\theta}\frac{y^2}{2}\Big|_0^\theta = \frac{\theta}{2}$$

4). Is the maximum likelihood estimator biased?

Since $E\big(\hat{\theta_{ml}} \mid \theta\big) = \frac{\theta}{2} \neq \theta$, it's a biased estimator.

5). For the more general case that $n \geq 1$, what is the expectation of $\hat{\theta}_{ml}$?

When $n \geq 1$, the expectation of MLE is $E\big(\hat{\theta_{ml}} \mid \theta\big) = \int_0^\theta f(y_{max} \mid \theta) y_{max} dy = \int_0^\theta \frac{1}{\theta} y dy = \frac{\theta}{2}$

6). Is the maximum likelihood estimator consistent?

the bias of the MLE in 5) is $\frac{\theta}{2}$ for any sample size, thus it's not consistent.

Note: the unbiased estimation of θ can be obtained by using moment estimator:

$$E(Y) = \int_0^\theta y \frac{1}{\theta} dy = \frac{\theta}{2} = \bar{y} = \frac{1}{n}\sum_i y_i \rightarrow \hat{\theta} = \frac{2}{n}\sum_i y_i$$

Question 4. Classical Linear Model 1

Background

The file WageData2.csv contains a dataset that has been used to quantify the impact of education on wage. One of the reasons we are proving another wage-equation exercise is that this area by far has the most (and most well-known) applications of instrumental variable techniques, the endogenity problem is obvious in this context, and the datasets are easy to obtain.

The Data

You are given a sample of 1000 individuals with their wage, education level, age, working experience, race (as an indicator), father's and mother's education level, whether the person lived in a rural area, whether the person lived in a city, IQ score, and two potential instruments, called $z1$ and $z2$.

The dependent variable of interest is $wage$ (or its transformation), and we are interested in measuring "return" to education, where return is measured in the increase (hopefully) in wage with an additional year of education.

Question 4.1

Conduct an univariate analysis (using tables, graphs, and descriptive statistics found in the last 7 lectures) of all of the variables in the dataset.

Also, create two variables: (1) natural log of wage (name it $logWage$) (2) square of experience (name it $experienceSquare$)

```
# load packages
library(car)
library(ggplot2)
library(lattice)
library(car)
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(AER)
```

```
## Loading required package: survival
## Loading required package: splines
```

```
library(ivpack)
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
# load data
setwd('~/GitHub/MIDS/MIDS-W271/lab2')
data <- read.csv("WageData2.csv", header = TRUE)
str(data)
```

```
## 'data.frame':    1000 obs. of  14 variables:
##  $ X            : int  191 2059 2072 945 1920 1927 1481 2571 437 1265 ...
##  $ wage         : int  951 288 509 647 225 454 565 479 615 641 ...
##  $ education    : int  12 8 12 18 10 10 12 13 16 12 ...
##  $ experience   : int  10 11 6 5 11 11 10 15 7 16 ...
##  $ age          : int  28 25 24 29 27 27 28 34 29 34 ...
##  $ raceColor    : int  0 1 0 0 1 1 1 0 0 0 ...
##  $ dad_education: int  NA NA 12 12 5 NA NA 7 12 4 ...
##  $ mom_education: int  12 7 9 12 5 1 NA 12 12 8 ...
##  $ rural        : int  0 1 1 0 1 1 1 1 0 0 ...
##  $ city         : int  1 0 1 1 0 0 1 1 1 0 ...
##  $ z1           : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ z2           : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ IQscore      : int  122 NA 127 110 NA NA NA NA 113 92 ...
##  $ logWage      : num  6.86 5.66 6.23 6.47 5.42 ...
```

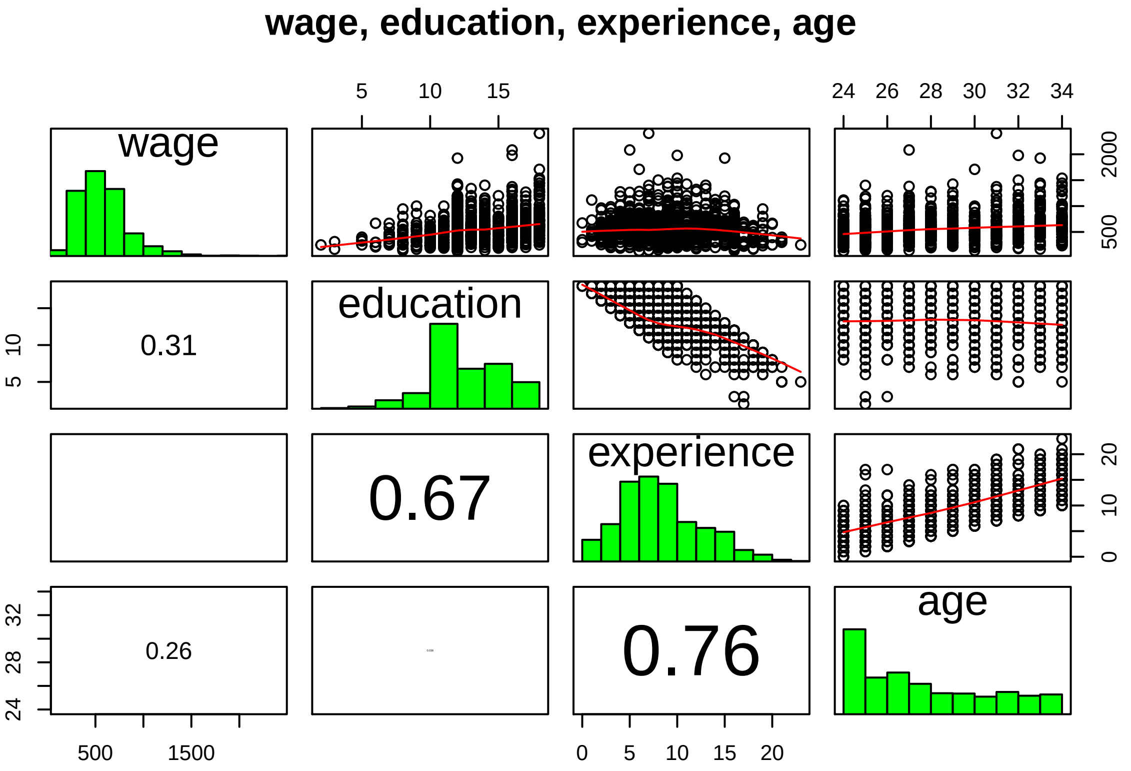
```
# show simple univariate stats for each variable
summary(data)
```

```
##        X               wage          education       experience
##  Min.   :   5.0   Min.   : 127.0   Min.   : 2.00   Min.   : 0.000
##  1st Qu.: 715.5   1st Qu.: 400.0   1st Qu.:12.00   1st Qu.: 6.000
##  Median :1431.5   Median : 543.0   Median :12.00   Median : 8.000
##  Mean   :1466.7   Mean   : 578.8   Mean   :13.22   Mean   : 8.788
##  3rd Qu.:2212.0   3rd Qu.: 702.5   3rd Qu.:16.00   3rd Qu.:11.000
##  Max.   :3009.0   Max.   :2404.0   Max.   :18.00   Max.   :23.000
##
##       age          raceColor     dad_education   mom_education
##  Min.   :24.00   Min.   :0.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:25.00   1st Qu.:0.000   1st Qu.: 8.00   1st Qu.: 8.00
##  Median :27.00   Median :0.000   Median :11.00   Median :12.00
##  Mean   :28.01   Mean   :0.238   Mean   :10.18   Mean   :10.45
##  3rd Qu.:30.00   3rd Qu.:0.000   3rd Qu.:12.00   3rd Qu.:12.00
##  Max.   :34.00   Max.   :1.000   Max.   :18.00   Max.   :18.00
##                                  NA's   :239     NA's   :128
##      rural             city             z1             z2
##  Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000
##  Median :0.000   Median :1.000   Median :0.00   Median :1.000
##  Mean   :0.391   Mean   :0.712   Mean   :0.44   Mean   :0.686
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.000   Max.   :1.00   Max.   :1.000
```

```
##
##     IQscore          logWage
##  Min.   : 50.0   Min.   :4.844
##  1st Qu.: 93.0   1st Qu.:5.991
##  Median :103.0   Median :6.297
##  Mean   :102.3   Mean   :6.263
##  3rd Qu.:113.0   3rd Qu.:6.555
##  Max.   :144.0   Max.   :7.785
##  NA's   :316
```

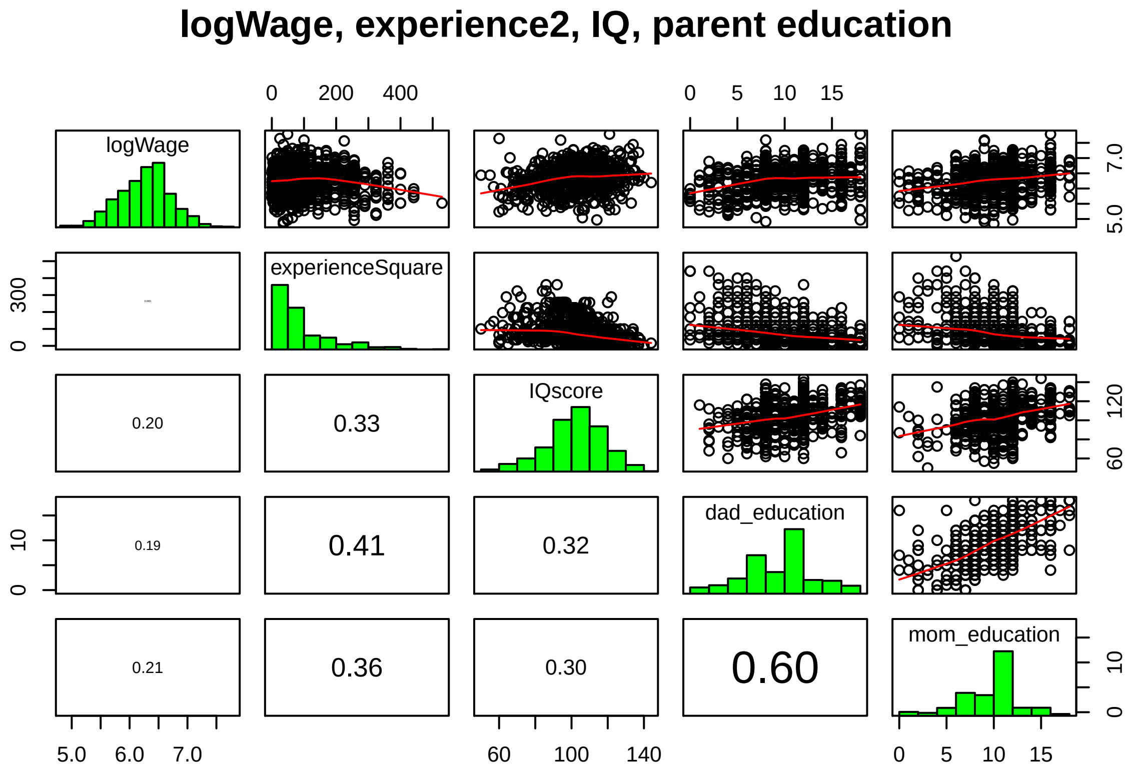
```
# create logWage and experienceSquare
data$logWage <- log(data$wage)
data$experienceSquare <- data$experience^2
```

Among the variables we are interested in for the analysis:



Question 4.2

Conduct a bivariate analysis (using tables, graphs, descriptive statistics found in the last 7 lectures) of *wage* and $logWage$ and all the other variables in the datasets.



Question 4.3

Regress $log(wage)$ on education, experience, age, and raceColor.

```
m4.3 <- lm(logWage~education+experience+age+raceColor, data=data)
```

1). Report all the estimated coefficients, their standard errors, t-statistics, F-statistic of the regression, R^2, $adjusted\ R^2$, and degrees of freedom.

```
summary(m4.3)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + age + raceColor,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35396 -0.25550  0.01074  0.24867  1.22932
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.961661   0.113346  43.774   <2e-16 ***
## education    0.079608   0.006376  12.486   <2e-16 ***
## experience   0.035372   0.003988   8.869   <2e-16 ***
## age                NA         NA      NA       NA
## raceColor   -0.260813   0.030453  -8.564   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3917 on 996 degrees of freedom
## Multiple R-squared:  0.236,  Adjusted R-squared:  0.2337
## F-statistic: 102.6 on 3 and 996 DF,  p-value: < 2.2e-16
```

2). Explain why the degrees of freedom takes on the specific value you observe in the regression output.

The degrees of freedom is 996, which is the number of observations minus number of predictors

3). Describe any unexpected results from your regression and how you would resolve them (if the intent is to estimate return to education, condition on race and experience).

Regression coefficient of age becomes NA –> collinearity?

To estimate return to education, condition on race and experience, we can add interaction terms between education, and experience and raceColor, to the model:

```r
m2 <- lm(logWage~education+education:experience+education:raceColor, data=data)
summary(m2)
```

```
##
## Call:
## lm(formula = logWage ~ education + education:experience + education:raceColor,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35989 -0.25381  0.01911  0.25382  1.24739
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.1924282  0.0789124  65.800  < 2e-16 ***
## education             0.0597973  0.0047078  12.702  < 2e-16 ***
## education:experience  0.0030623  0.0002979  10.279  < 2e-16 ***
## education:raceColor  -0.0185244  0.0024049  -7.703 3.21e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3898 on 996 degrees of freedom
## Multiple R-squared:  0.2434, Adjusted R-squared:  0.2411
## F-statistic: 106.8 on 3 and 996 DF,  p-value: < 2.2e-16
```

take derivative of $wage$ with respect to $education$, we obtain the conditional return to education on race and experience:

$$\frac{d(wage)}{d(education)} = 0.0597973 + 0.0030623 \times experience - 0.0185244 \times raceColor$$

4). Interpret the coefficient estimate associated with education

Hold all other factors constant, 1 more year of eduction will increase the wage by 8%.

5). Interpret the coefficient estimate associated with experience

Hold all other factors constant, 1 more year of experience will increase the wage by 3.5%.

Question 4.4

Regress $log(wage)$ on education, experience, experienceSquare, and raceColor.

```
m4.4 <- lm(logWage~education+experience+experienceSquare+raceColor, data=data)
summary(m4.4)
```

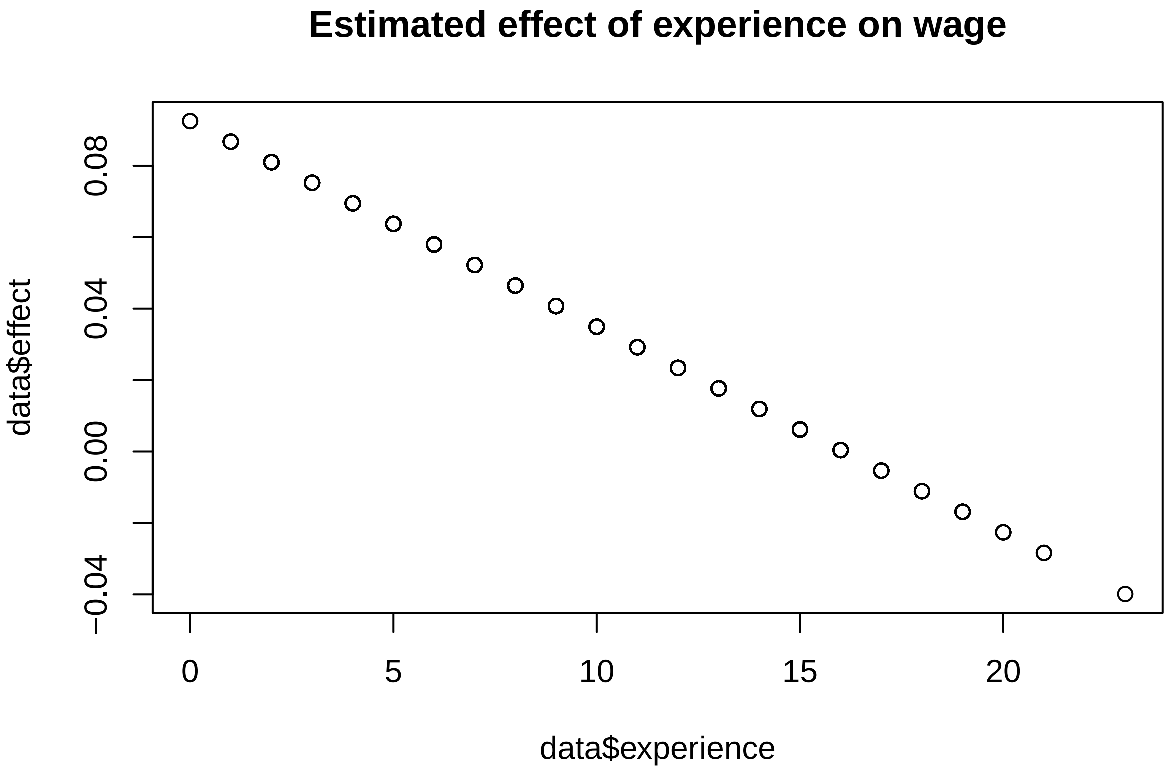
```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38464 -0.25558  0.01909  0.25782  1.24410
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.7355175  0.1197719  39.538  < 2e-16 ***
## education         0.0794641  0.0062917  12.630  < 2e-16 ***
## experience        0.0924930  0.0115147   8.033 2.68e-15 ***
## experienceSquare -0.0028779  0.0005452  -5.279 1.60e-07 ***
## raceColor        -0.2627226  0.0300528  -8.742  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3865 on 995 degrees of freedom
## Multiple R-squared:  0.2569, Adjusted R-squared:  0.2539
## F-statistic: 85.98 on 4 and 995 DF,  p-value: < 2.2e-16
```

1). Plot a graph of the estimated effect of experience on wage.

Take derivative of $wage$ with respect to $experience$:

$$\frac{d(logWage)}{d(experience)} = 0.092493 - 0.005756 \times experience \tag{4.4}$$

```
data$effect <- 0.092493-0.005756*data$experience
plot(data$experience, data$effect, main = 'Estimated effect of experience on wage')
```



2). What is the estimated effect of experience on wage when experience is 10 years?

when experience is 10 years, plug in equation 4.4, we have the effect of 0.034933.

the more experience one has, the smaller effect on the wage increase.

Question 4.5

Regress *logWage* on education, experience, experienceSquare, raceColor, dad_education, mom_education, rural, city.

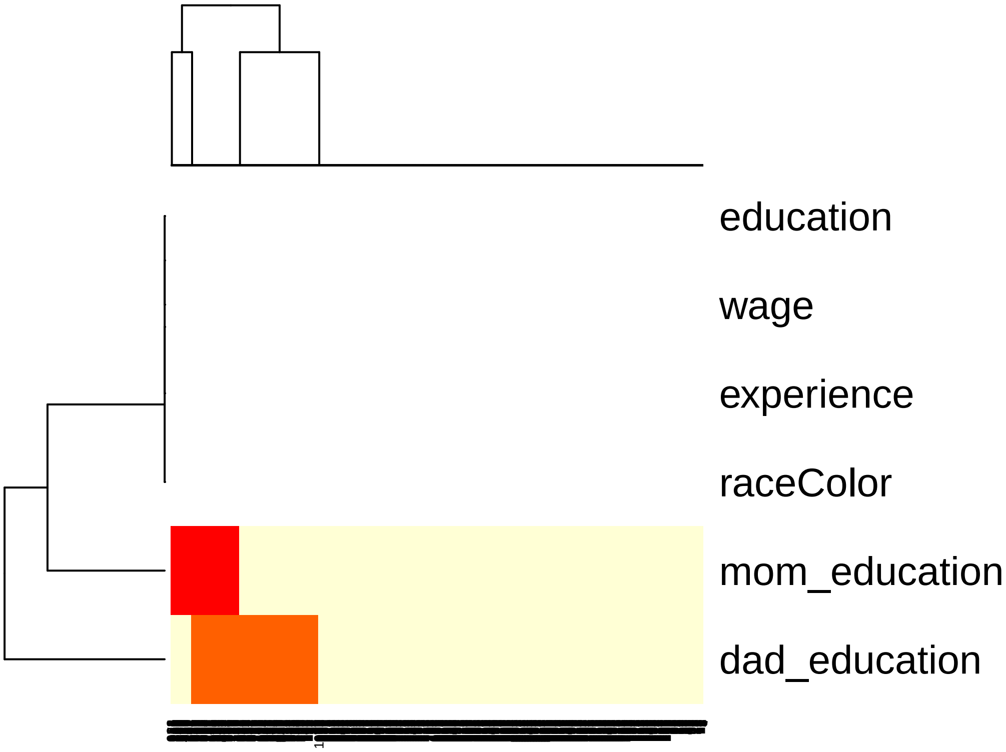
```
m4.5 <- lm(logWage ~ education+experience+experienceSquare+raceColor
           +dad_education+mom_education+rural+city, data=data)
summary(m4.5)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education + mom_education + rural + city,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2961 -0.2240  0.0160  0.2454  1.0404
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.6422296  0.1408825  32.951  < 2e-16 ***
## education         0.0681701  0.0077409   8.806  < 2e-16 ***
## experience        0.0973419  0.0133133   7.312  7.1e-13 ***
## experienceSquare -0.0029568  0.0006678  -4.428  1.1e-05 ***
## raceColor        -0.2130226  0.0425014  -5.012  6.8e-07 ***
## dad_education    -0.0011474  0.0050988  -0.225  0.82202
## mom_education     0.0113176  0.0061886   1.829  0.06785 .
## rural            -0.0919377  0.0314151  -2.927  0.00354 **
## city              0.1782137  0.0323826   5.503  5.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3786 on 714 degrees of freedom
##   (277 observations deleted due to missingness)
## Multiple R-squared:  0.2746, Adjusted R-squared:  0.2665
## F-statistic: 33.79 on 8 and 714 DF,  p-value: < 2.2e-16
```

1). What are the number of observations used in this regression? Are missing values a problem? Analyze the missing values, if any, and see if there is any discernible pattern with wage, education, experience, and raceColor.

The number of observation used in this regress is `723`.

```
na.index=is.na(data[,c('mom_education','dad_education','wage','education',
                      'experience','raceColor')])
na.index[na.index]=1
heatmap(1-t(na.index))
```

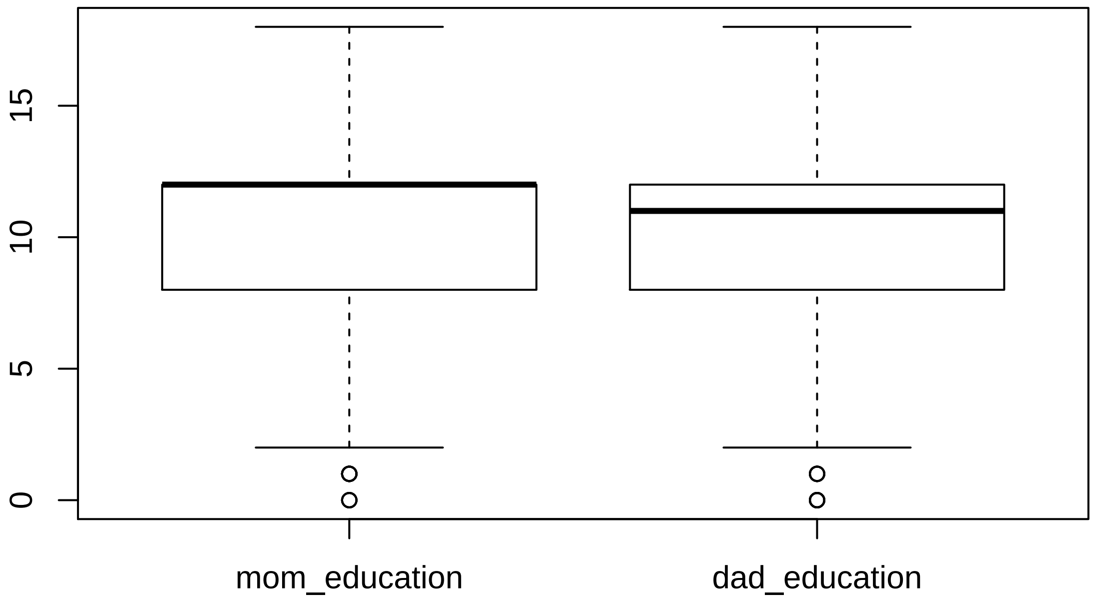


There are `128` missing value in mom_education and `239` in dad_education. All other variables have complete records. The distribution of the missing values, with respect to the rows and variables, can be seen in the heatmap as red and orange blocks.

2). Do you just want to "throw away" these observations?

It is not a good strategy to "throw away" observations with missing education of mom and/or dad education, doing so we lose a lot useful infomration in other variables. In addition, from the boxplot below we can see the variation of these two variables is not quite large, we would apply some rule to interpolate the missing values.

```
boxplot(data[,c('mom_education','dad_education')])
```



3). How about blindly replace all of the missing values with the average of the observed values of the corresponding variable? Rerun the original regression using all of the observations?

```
# replace mom/dad education with their means respectively
mom_educ_mean <- mean(data$mom_education, na.rm=T)
dad_educ_mean <- mean(data$dad_education, na.rm=T)
m_miss = is.na(data$mom_education)
d_miss = is.na(data$dad_education)
data$dad_education_fill = data$dad_education
data$mom_education_fill = data$mom_education
data$mom_education_fill[m_miss] = mom_educ_mean
data$dad_education_fill[d_miss] = dad_educ_mean
# rerun the regression
m4.5.3 <- lm(logWage ~ education+experience+experienceSquare+raceColor
             +dad_education_fill+mom_education_fill+rural+city, data=data)
summary(m4.5.3)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education_fill + mom_education_fill + rural +
##     city, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30741 -0.23286  0.01943  0.24786  1.28807
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.729e+00  1.226e-01  38.584  < 2e-16 ***
## education            7.097e-02  6.499e-03  10.920  < 2e-16 ***
## experience           8.958e-02  1.124e-02   7.970 4.36e-15 ***
## experienceSquare    -2.678e-03  5.318e-04  -5.036 5.65e-07 ***
## raceColor           -2.313e-01  3.099e-02  -7.464 1.84e-13 ***
## dad_education_fill -3.513e-05  4.416e-03  -0.008 0.993656
## mom_education_fill  3.485e-03  5.009e-03   0.696 0.486742
## rural               -9.529e-02  2.638e-02  -3.612 0.000319 ***
## city                 1.671e-01  2.703e-02   6.183 9.21e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3764 on 991 degrees of freedom
## Multiple R-squared:  0.2981, Adjusted R-squared:  0.2925
## F-statistic: 52.62 on 8 and 991 DF,  p-value: < 2.2e-16
```

The new model is similar with previous one built with limited observations, in terms of the effect of mom/dad education on one's wage. However, the effect of education and experience both have changed.

4). How about regress the variable(s) with missing values on education, experience, and raceColor, and use this regression(s) to predict (i.e. "impute") the missing values and then rerun the original regression using all of the observations?

```r
# regression for parent education
m4.5.4.mom <- lm(mom_education~education+experience+raceColor, data=data)
m4.5.4.dad <- lm(dad_education~education+experience+raceColor, data=data)
# predict for missing records
m_miss = is.na(data$mom_education)
d_miss = is.na(data$dad_education)
fill_mom <- predict.lm(m4.5.4.mom, data[m_miss,])
fill_dad <- predict.lm(m4.5.4.dad, data[d_miss,])
# fill back in
data$dad_education_fill = data$dad_education
data$mom_education_fill = data$mom_education
data$mom_education_fill[m_miss] = fill_mom
data$dad_education_fill[d_miss] = fill_dad
# rerun the regression
m4.5.4 <- lm(logWage ~ education+experience+experienceSquare+raceColor
             +dad_education_fill+mom_education_fill+rural+city, data=data)
summary(m4.5.4)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education_fill + mom_education_fill + rural +
##     city, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30581 -0.22943  0.01788  0.24773  1.28261
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.726839   0.121266  38.979  < 2e-16 ***
## education            0.070083   0.006690  10.476  < 2e-16 ***
## experience           0.089482   0.011223   7.973 4.24e-15 ***
## experienceSquare    -0.002654   0.000532  -4.990 7.14e-07 ***
## raceColor           -0.226516   0.032067  -7.064 3.05e-12 ***
## dad_education_fill   0.002360   0.004739   0.498 0.618562
## mom_education_fill   0.002401   0.005170   0.464 0.642407
## rural               -0.094838   0.026396  -3.593 0.000343 ***
## city                 0.166532   0.027054   6.155 1.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3764 on 991 degrees of freedom
## Multiple R-squared:  0.2983, Adjusted R-squared:  0.2926
## F-statistic: 52.65 on 8 and 991 DF,  p-value: < 2.2e-16
```

with the imputed values for mom/dad education, the effect of education and experience remain largely the same with previous one with data filled by mean value.

5). Compare the results of all of these regressions. Which one, if at all, would you prefer?

```
summary(m4.5.4.mom)
```

```
##
## Call:
## lm(formula = mom_education ~ education + experience + raceColor,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.552  -1.330   0.216   1.747   7.215
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.59262    0.82675   6.765 2.46e-11 ***
## education    0.43314    0.04636   9.342  < 2e-16 ***
## experience  -0.07676    0.02981  -2.575   0.0102 *
## raceColor   -1.46754    0.23241  -6.315 4.32e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.669 on 868 degrees of freedom
##   (128 observations deleted due to missingness)
## Multiple R-squared:  0.2736, Adjusted R-squared:  0.2711
## F-statistic:   109 on 3 and 868 DF,  p-value: < 2.2e-16
```

```
summary(m4.5.4.dad)
```

```
##
## Call:
## lm(formula = dad_education ~ education + experience + raceColor,
##     data = data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0912  -1.9700   0.0488   2.0567   9.3408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.93928    1.01939   4.845 1.53e-06 ***
## education    0.50248    0.05748   8.741  < 2e-16 ***
## experience  -0.14796    0.03662  -4.041 5.88e-05 ***
## raceColor   -2.12117    0.31189  -6.801 2.11e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.122 on 757 degrees of freedom
##   (239 observations deleted due to missingness)
## Multiple R-squared:  0.309,  Adjusted R-squared:  0.3062
## F-statistic: 112.8 on 3 and 757 DF,  p-value: < 2.2e-16
```

from the regression model of parents education we can see the R^2 is modest, although all predictors are significant. Given that parent education variables are insignificant in either of the 3 models for one's wage, the imputed values may just bring more noise than useful information to the model. Thus I would prefer to omit these two variables and build a model with all 1000 observations on the rest of the variables.

Question 4.6

1). Consider using z_1 as the instrumental variable (IV) for education. What assumptions are needed on z_1 and the error term (call it, u)?

In order for z_1 to be an instrument variable, it should be highly correlated with our predictor(s) of interests (education, experience), while uncorrelated with the error term, namely $cov(z_1, u) = 0$

2). Suppose z_1 is an indicator representing whether or not an individual lives in an area in which there was a recent policy change to promote the importance of education. Could z_1 be correlated with other unobservables captured in the error term?

Although z_1 here presumably will be correlated with one's education, tt is possible that z_1 becomes correlated with other unobservables in the error term. For example, with the policy change to promote the importance of education in an areea, more high income families are intended to move to the area, and family income/wealthness could be a factor that correlated with one's own wage. Thus with z_1 as instrument variable, the measured effect of education on wage could be a hetergenerous effect for certain groups of people.

3). Using the same specification as that in question 4.5, estimate the equation by 2SLS, using both z_1 and z_2 as instrument variables. Interpret the results. How does the coefficient estimate on education change?

```
ols1 <- lm(logWage ~ education + experience, data=data)
se_ols1 <- robust.se(ols1)[,2]
```

```
## [1] "Robust Standard Errors"
```

```
ols2 <- lm(logWage ~ education + experience + experienceSquare + raceColor, data=data)
se_ols2 <- robust.se(ols2)[,2]
```

```
## [1] "Robust Standard Errors"
```

```
ols3 <- lm(logWage ~ education + experience + experienceSquare + raceColor
           + dad_education + mom_education, data=data)
se_ols3 <- robust.se(ols3)[,2]
```

```
## [1] "Robust Standard Errors"
```

```
ols4 <- lm(logWage ~ education + experience + experienceSquare + raceColor
           + dad_education + mom_education + rural + city, data=data)
se_ols4 <- robust.se(ols4)[,2]
```

```
## [1] "Robust Standard Errors"
```

```
tsls1 <- ivreg(logWage ~ education + experience | factor(z1)*factor(z2) + experience,
               data = data)
se_tsls1 <- robust.se(tsls1)[,2]
```

```
## [1] "Robust Standard Errors"
```

```
tsls2 <- ivreg(logWage ~ education + experience + experienceSquare + raceColor |
                 factor(z1)*factor(z2) + experience + experienceSquare, data=data)
se_tsls2 <- robust.se(tsls2)[,2]
```

```
## [1] "Robust Standard Errors"
```

```
tsls3 <- ivreg(logWage ~ education + experience + experienceSquare + raceColor
               + dad_education + mom_education | factor(z1)*factor(z2) + experience
               + experienceSquare + dad_education + mom_education, data=data)
se_tsls3 <- robust.se(tsls3)[,2]
```

```
## [1] "Robust Standard Errors"
```

```
tsls4 <- ivreg(logWage ~ education + experience + experienceSquare + raceColor
               + dad_education + mom_education + rural + city | factor(z1)*factor(z2)
               + experience + experienceSquare + dad_education + mom_education
               + rural + city, data=data)
se_tsls4 <- robust.se(tsls4)[,2]
```

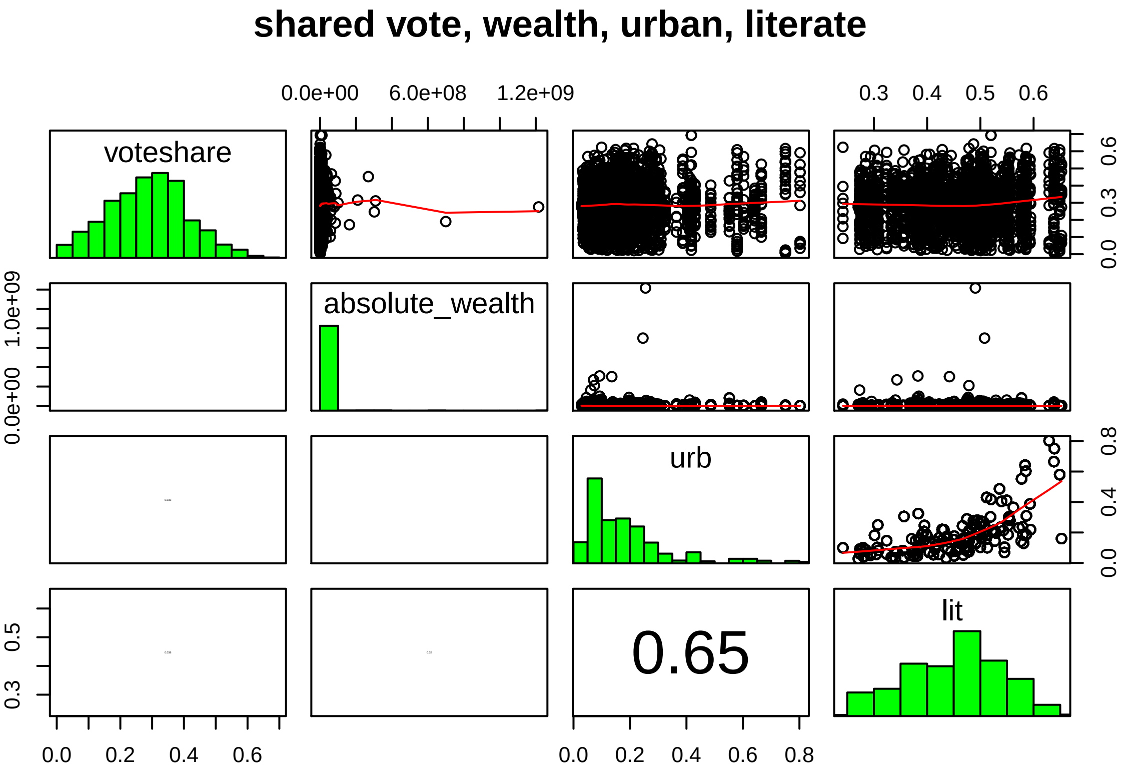
```
## [1] "Robust Standard Errors"
```

```
# generate regression table
#stargazer(ols1, tsls1, ols2, tsls2, ols3, tsls3, ols4, tsls4,
#          se = list(se_ols1, se_tsls1, se_ols2, se_tsls2,
#                    se_ols3, se_tsls3, se_ols4, se_tsls4),
#          covariate.labels=c("education", "experience", "experience squared",
#                             "race (1 = black)", "dad Education", "mom education"),
#          dep.var.labels = "Log Weekly Wage",
#          omit = c("city*","rural"),
#          out = "Q4_table.html", df= F,
#          omit.labels = c("rural", "city")
#          )
```

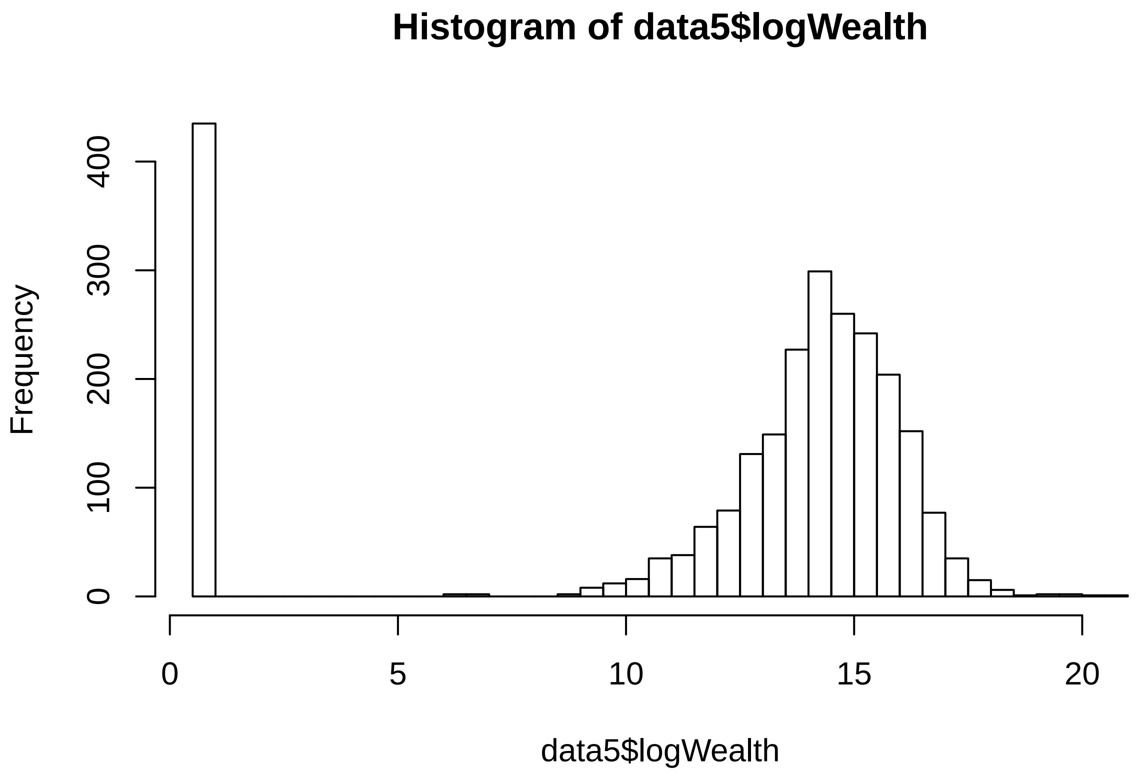
Question 5. Classical Linear Model 2

The dataset, ”wealthy candidates.csv”, contains candidate level electoral data from a developing country. Politically, each region (which is a subset of the country) is divided in to smaller electoral districts where the candidate with the most votes wins the seat. This dataset has data on the financial wealth and electoral performance (voteshare) of electoral candidates. We are interested in understanding whether or not wealth is an electoral advantage. In other words, do wealthy candidates fare better in elections than their less wealthy peers?

```
## 'data.frame':    2498 obs. of  6 variables:
##  $ X              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ region         : Factor w/ 3 levels "Region 1","Region 2",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ urb            : num  0.1491 0.1491 0.0918 0.1017 0.0614 ...
##  $ lit            : num  0.428 0.428 0.458 0.306 0.273 ...
##  $ voteshare      : num  0.417 0.114 0.298 0.484 0.311 ...
##  $ absolute_wealth: num  5110593 100000 55340 207000 1307408 ...
```



Upon further investigation we find the distribution of wealth is highly skewed. To mitigate the nonnormality here, we take the log of absolute_wealth. In addition there are 435 candidates who has an absolute wealth of $2, these may be due to some cutoff line in data collection, where anything below would be $2. It’s probably a good idea to set it to $1, so that after taking log it becomes zero, and this group becomes baseline.



We can see after taking log the distribution of wealth variable is not as heavily skewed as previously, except a outlier cluster with value $log2.

1). Begin with a parsimonious, yet appropriate, specification. Why did you choose this model? Are your results statistically significant? Based on these results, how would you answer the research question? Is there a linear relationship between wealth and electoral performance?

Let's simply assume there is a positive correlation between log candidate wealth and electoral performance, and build a linear model:

```
m5 <- lm(voteshare ~ logWealth, data = data5)
summary(m5)
```

```
##
## Call:
## lm(formula = voteshare ~ logWealth, data = data5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28526 -0.08796  0.00551  0.07933  0.40256
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.2724581  0.0060004  45.407  < 2e-16 ***
## logWealth   0.0012924  0.0004574   2.825  0.00476 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1231 on 2495 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.003189,   Adjusted R-squared:  0.00279
## F-statistic: 7.983 on 1 and 2495 DF,  p-value: 0.00476
```

and the model indicates financial wealth has statistical significant effect on electoral performance, but the practical effect size is quite small.

2). A team-member suggests adding a quadratic term to your regression. Based on your prior model, is such an addition warranted? Add this term and interpret the results. Do wealthier candidates fare better in elections?

A quadratic term is added when the treatment effect is not constant as the predictor value changes. Here by taking derivative with respect to the predictor, we obtain the effect on the change rate shared vote attributable to wealth.

```
m5 <- lm(voteshare ~ logWealth + I(logWealth**2), data = data5)
summary(m5)
```

```
##
## Call:
## lm(formula = voteshare ~ logWealth + I(logWealth^2), data = data5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28475 -0.08803  0.00391  0.08011  0.40438
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.2779620  0.0066522  41.785   <2e-16 ***
## logWealth      -0.0028263  0.0022021  -1.283    0.199
## I(logWealth^2)  0.0002543  0.0001330   1.912    0.056 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1231 on 2494 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.004648,   Adjusted R-squared:  0.00385
## F-statistic: 5.824 on 2 and 2494 DF,  p-value: 0.002997
```

here the effect is not significant.

3). Another team member suggests that it is important to take into account the fact that different regions have different electoral contexts. In particular, the relationship between candidate wealth and electoral performance might be different across states. Modify your model and report your results. Test the hypothesis that this addition is not needed.

We evaluate the interaction terms between logWage and region:

```
m5 <- lm(voteshare ~ logWealth*factor(region), data = data5)
summary(m5)
```

```
##
## Call:
## lm(formula = voteshare ~ logWealth * factor(region), data = data5)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31125 -0.08569  0.00857  0.07952  0.39530
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.2757166  0.0067558  40.812  < 2e-16
## logWealth                       -0.0002482  0.0005272  -0.471 0.637779
## factor(region)Region 2          -0.0436968  0.0169807  -2.573 0.010130
## factor(region)Region 3          -0.0078181  0.0229382  -0.341 0.733258
## logWealth:factor(region)Region 2  0.0049088  0.0012780   3.841 0.000126
## logWealth:factor(region)Region 3  0.0036368  0.0017203   2.114 0.034607
##
## (Intercept)                     ***
## logWealth
## factor(region)Region 2          *
## factor(region)Region 3
## logWealth:factor(region)Region 2 ***
## logWealth:factor(region)Region 3 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1219 on 2491 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.02486,    Adjusted R-squared:  0.0229
## F-statistic:  12.7 on 5 and 2491 DF,  p-value: 3.295e-12
```

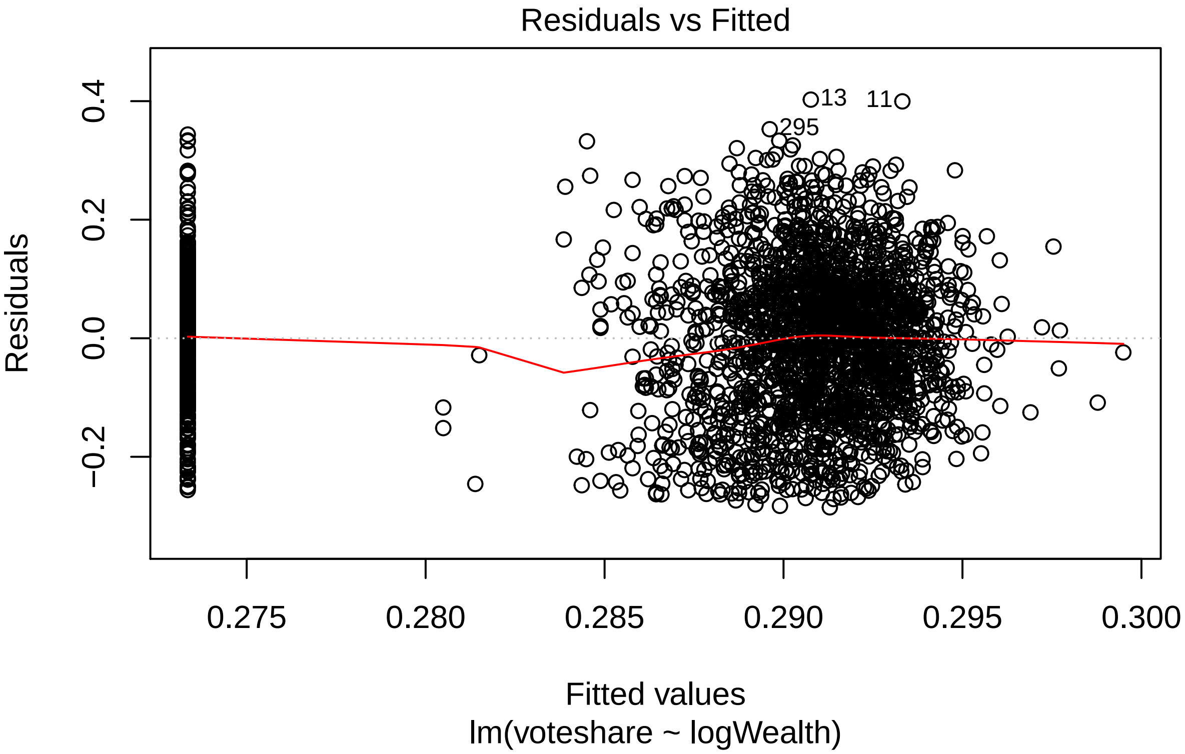
```r
coeftest(m5, vcov = vcovHC)
```

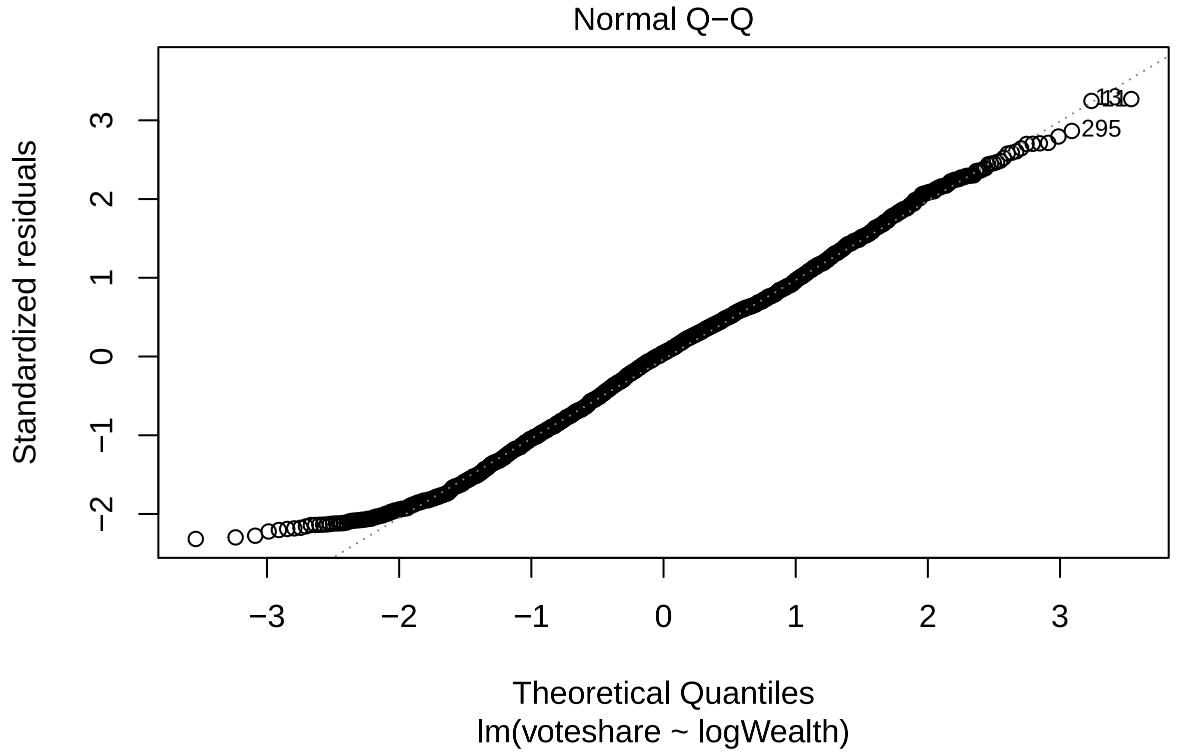
```
##
## t test of coefficients:
##
##                                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)                      0.27571655  0.00575284 47.9271 < 2.2e-16
## logWealth                       -0.00024823  0.00043760 -0.5672 0.5706006
## factor(region)Region 2          -0.04369681  0.01719474 -2.5413 0.0111046
## factor(region)Region 3          -0.00781812  0.03278908 -0.2384 0.8115621
## logWealth:factor(region)Region 2  0.00490878  0.00127972  3.8358 0.0001283
## logWealth:factor(region)Region 3  0.00363684  0.00240211  1.5140 0.1301475
##
## (Intercept)                     ***
## logWealth
## factor(region)Region 2          *
## factor(region)Region 3
## logWealth:factor(region)Region 2 ***
## logWealth:factor(region)Region 3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

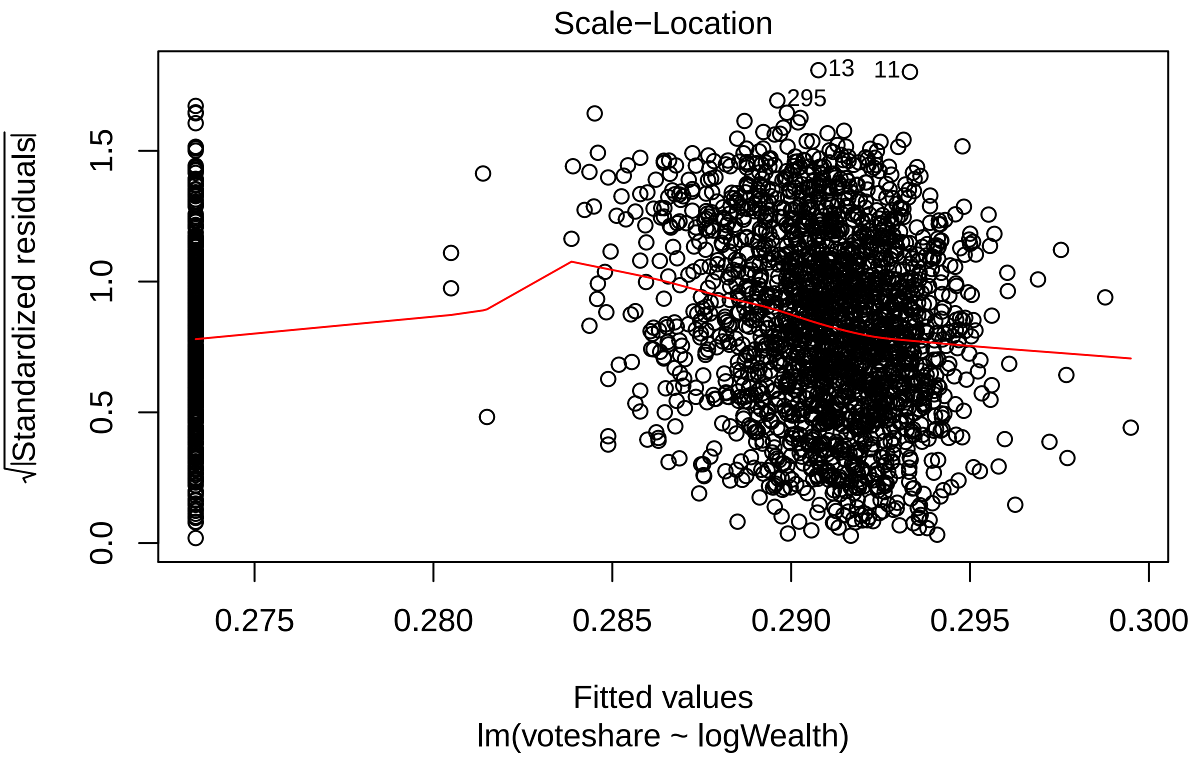
the model indicates that being in region 2 and 3 the effect of wealth on electoral performance is significant. And with heteroschedasticy considered, the test shows being in region 2 the wealth effect is significant.

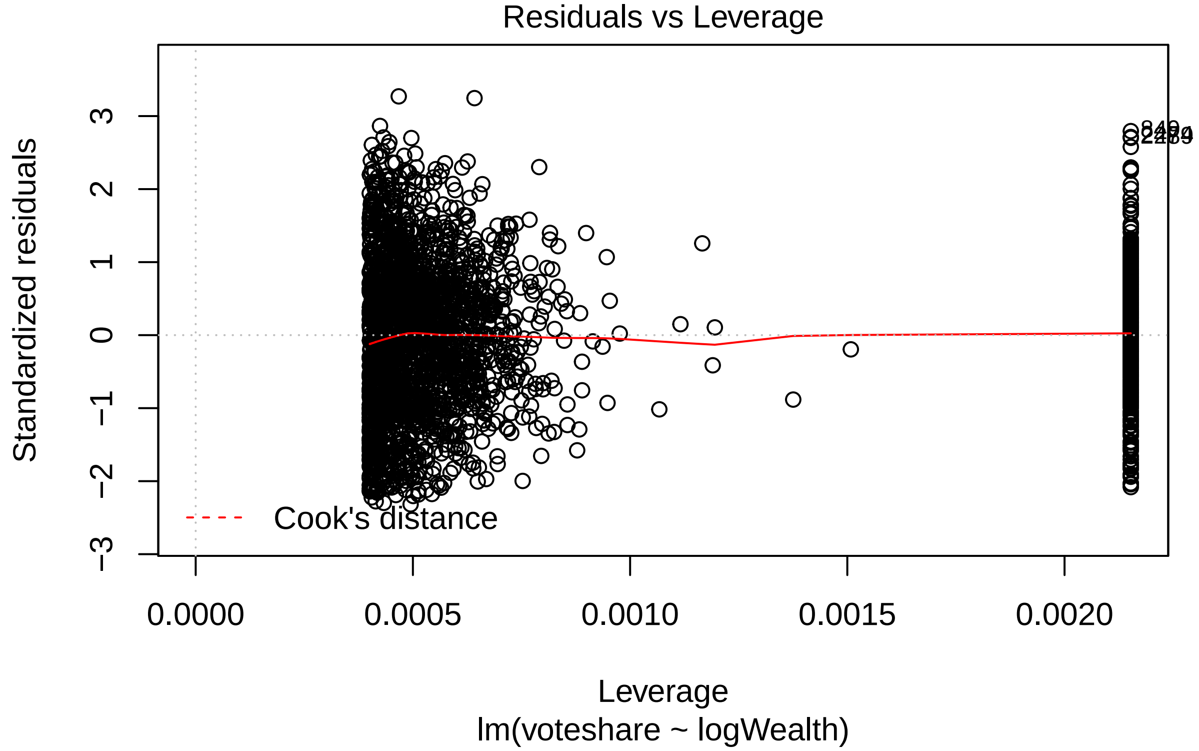
4). Return to your parsimonious model. Do you think you have found a causal and unbiased estimate? Please state the conditions under which you would have an unbiased and causal estimates. Do these conditions hold?

We evaluate the diagnostic plot of the parsimonious model.









5). Someone proposes a difference in difference design. Please write the equation for such a model. Under what circumstances would this design yield a causal effect?

$$cvoteshare = \beta_0 + \beta_1 cwealth + cu_i$$

given $cov(\Delta u_i, \Delta wealth) = 0$

Question 6. Classical Linear Model 3

Your analytics team has been tasked with analyzing aggregate revenue, cost and sales data, which have been provided to you in the R workspace/data frame retailSales.Rdata.

Your task is two fold. First, your team is to develop a model for predicting (forecasting) revenues. Part of the model development documentation is a backtesting exercise where you train your model using data from the first two years and evaluate the model's forecasts using the last two years of data.

Second, management is equally interested in understanding variables that might affect revenues in support of management adjustments to operations and revenue forecasts. You are also to identify factors that affect revenues, and discuss how useful management's planned revenue is for forecasting revenues.

Your analysis should address the following:

) Exploratory Data Analysis: focus on bivariate and multivariate relationships) Be sure to assess conditions and identify unusual observations *) Is the change in the average revenue different from 95 cents when the planned revenue increases by $1?*) Explain what interaction terms in your model mean in context supported by data visualizations *) Give two reasons why the OLS model coefficients may be biased and/or not consistent, be specific.*) Propose (but do not actually implement) a plan for an IV approach to improve your forecasting model.