Lab 2

Ron Cordell, Lei Yang, Subhashini Raghunathan

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 $experienceSquare$)

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.2.3
```

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

```
library(car)
library(sandwich)
setwd("C:/Subha/WS271-Regression/Labs/lab2_w271_2016Spring")
wd = read.csv("WageData2.csv")
str(wd)
```

```
## '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 ...
```

```
attach(wd)
```

Dataset has 1000 observations

Wage: ranges from about 100 to 2500 with a mean of about 580 (units not clear) Positively skewed, no missing values

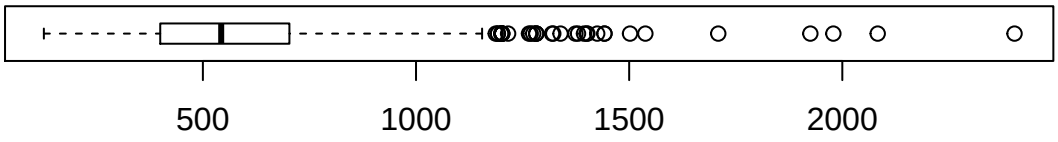
```
summary(wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   127.0   400.0   543.0   578.8   702.5  2404.0
```

```
str(wage)
```

```
##  int [1:1000] 951 288 509 647 225 454 565 479 615 641 ...
```

```
nf <- layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,3))
par(mar=c(3.1, 3.1, 1.1, 2.1))
boxplot(wage, horizontal=TRUE,  outline=TRUE)
hist(wage)
```



Education: ranges from 2 to 18, unit must be years Negatively skewed

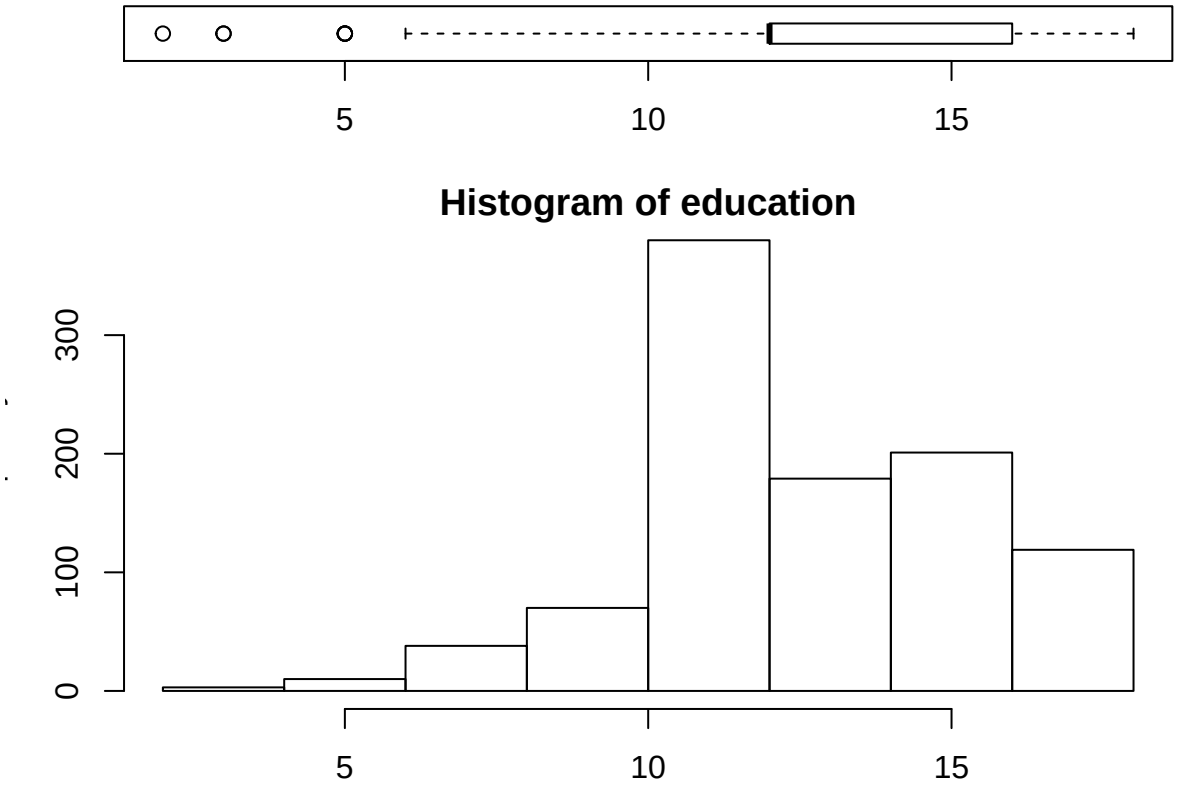
```
summary(education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   12.00   12.00   13.22   16.00   18.00
```

```
str(education)
```

```
##  int [1:1000] 12 8 12 18 10 10 12 13 16 12 ...
```

```
nf <- layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,3))
par(mar=c(3.1, 3.1, 1.1, 2.1))
boxplot(education, horizontal=TRUE,  outline=TRUE)
hist(education)
```



Experience: ranges from 0 to 23 years, mean = 8.8 Highly positivey skewed

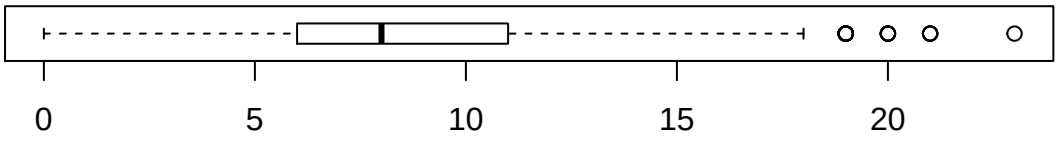
```
summary(experience)
```

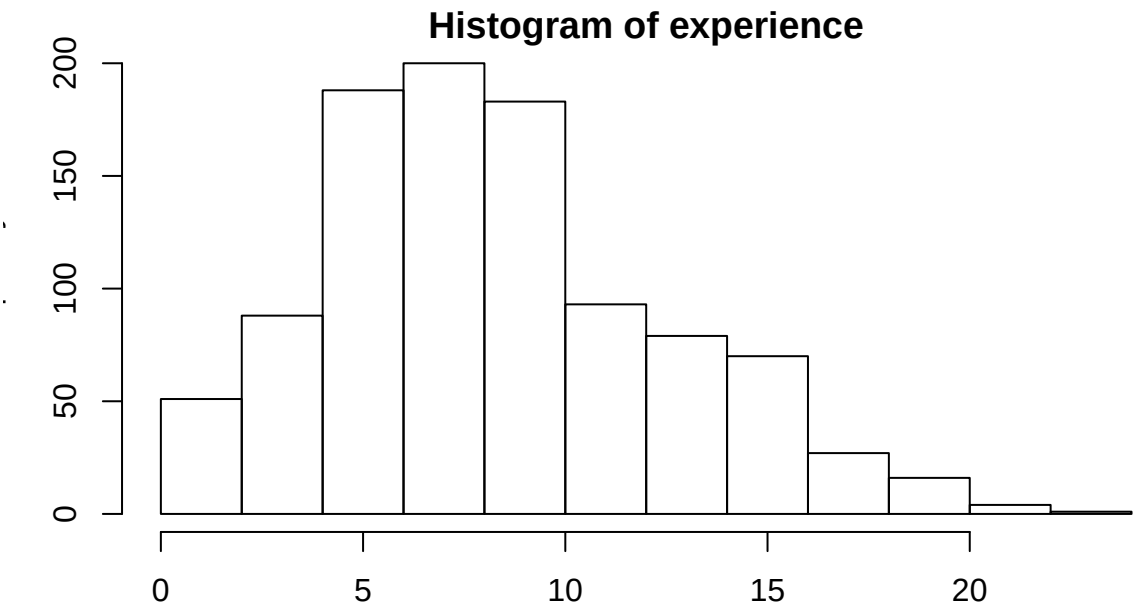
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.000   8.000   8.788  11.000  23.000
```

```
str(experience)
```

```
##  int [1:1000] 10 11 6 5 11 11 10 15 7 16 ...
```

```
nf <- layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,3))
par(mar=c(3.1, 3.1, 1.1, 2.1))
boxplot(experience, horizontal=TRUE,  outline=TRUE)
hist(experience)
```





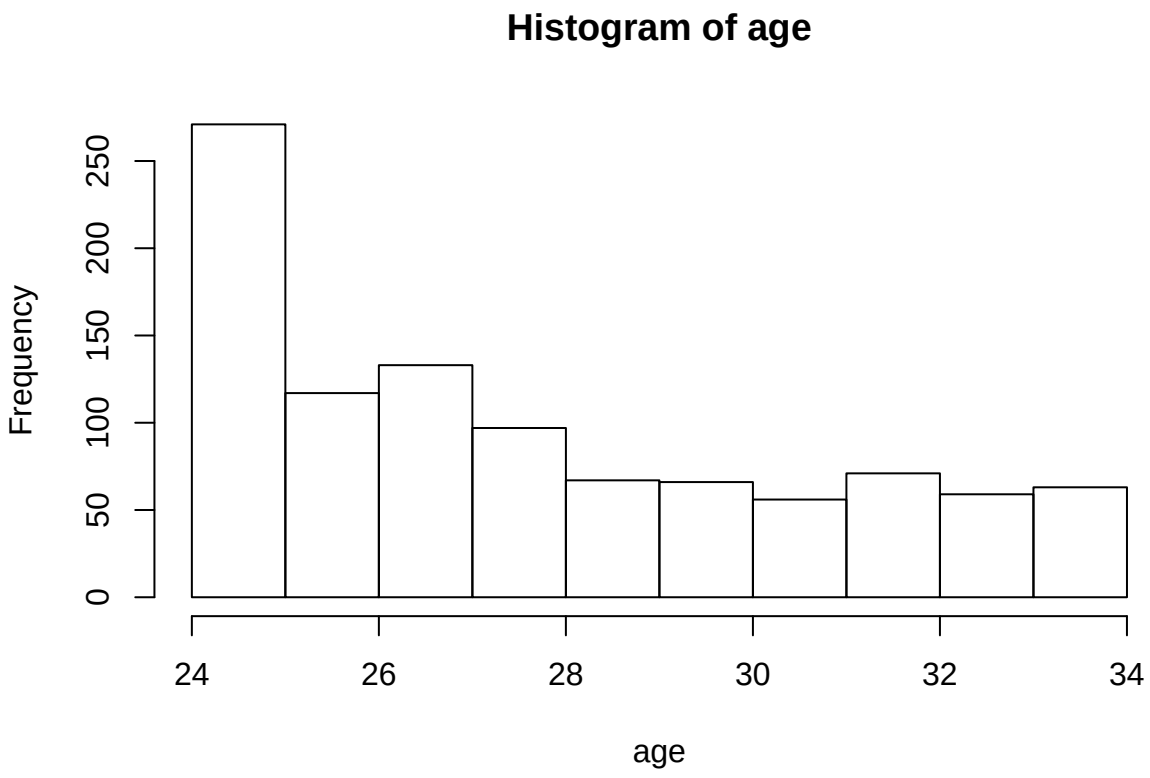

```
summary(age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.00   25.00   27.00   28.01   30.00   34.00
```

```
str(age)
```

```
##  int [1:1000] 28 25 24 29 27 27 28 34 29 34 ...
```

```
hist(age)
```



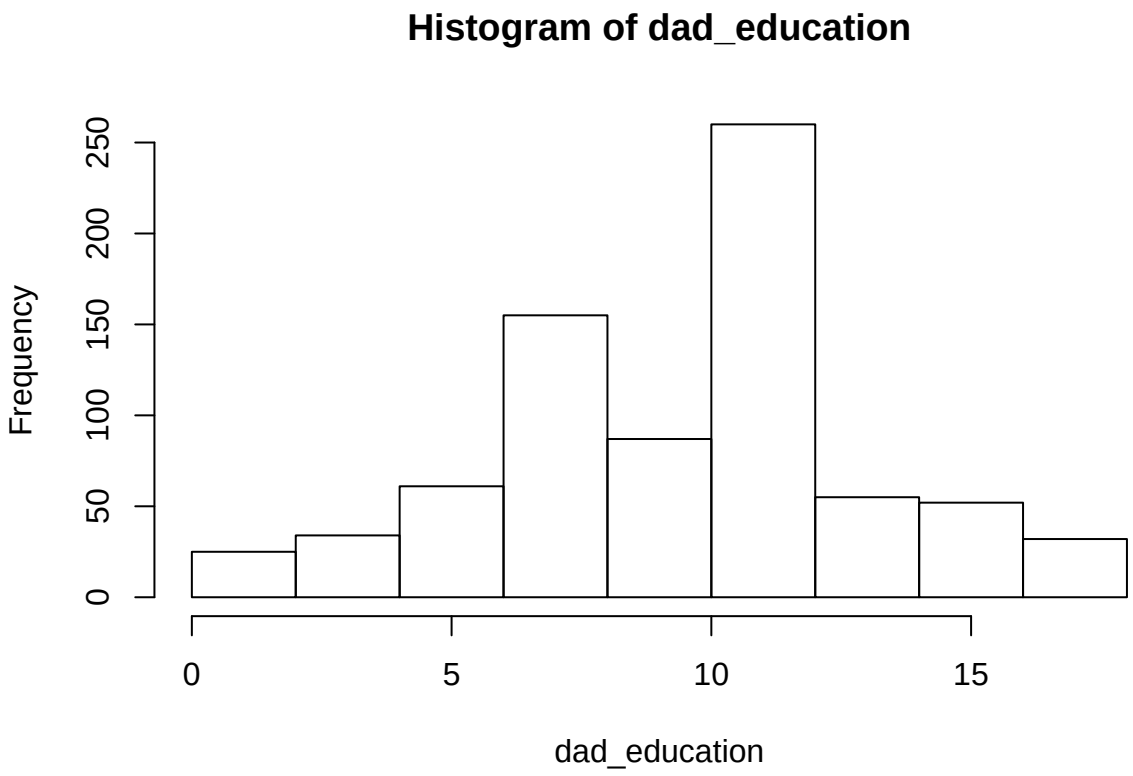

```
summary(dad_education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    8.00   11.00   10.18   12.00   18.00     239
```

```
str(dad_education)
```

```
##  int [1:1000] NA NA 12 12 5 NA NA 7 12 4 ...
```

```
hist(dad_education)
```




```
summary(mom_education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    8.00   12.00   10.45   12.00   18.00     128
```

```
str(mom_education)
```

```
##  int [1:1000] 12 7 9 12 5 1 NA 12 12 8 ...
```

```
hist(mom_education)
```

Has quite a few missing observations (316)

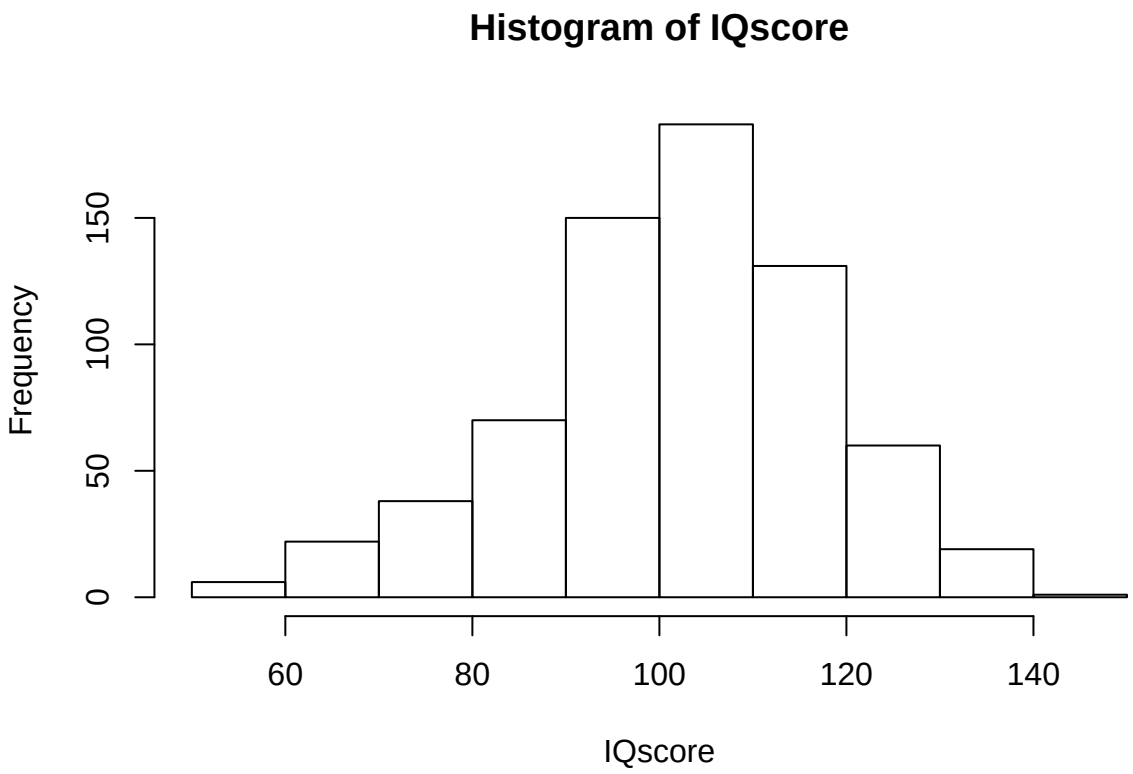
```
summary(IQscore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    50.0    93.0   103.0   102.3   113.0   144.0     316
```

```
str(IQscore)
```

```
##  int [1:1000] 122 NA 127 110 NA NA NA NA 113 92 ...
```

```
hist(IQscore)
```



almost normal distribution

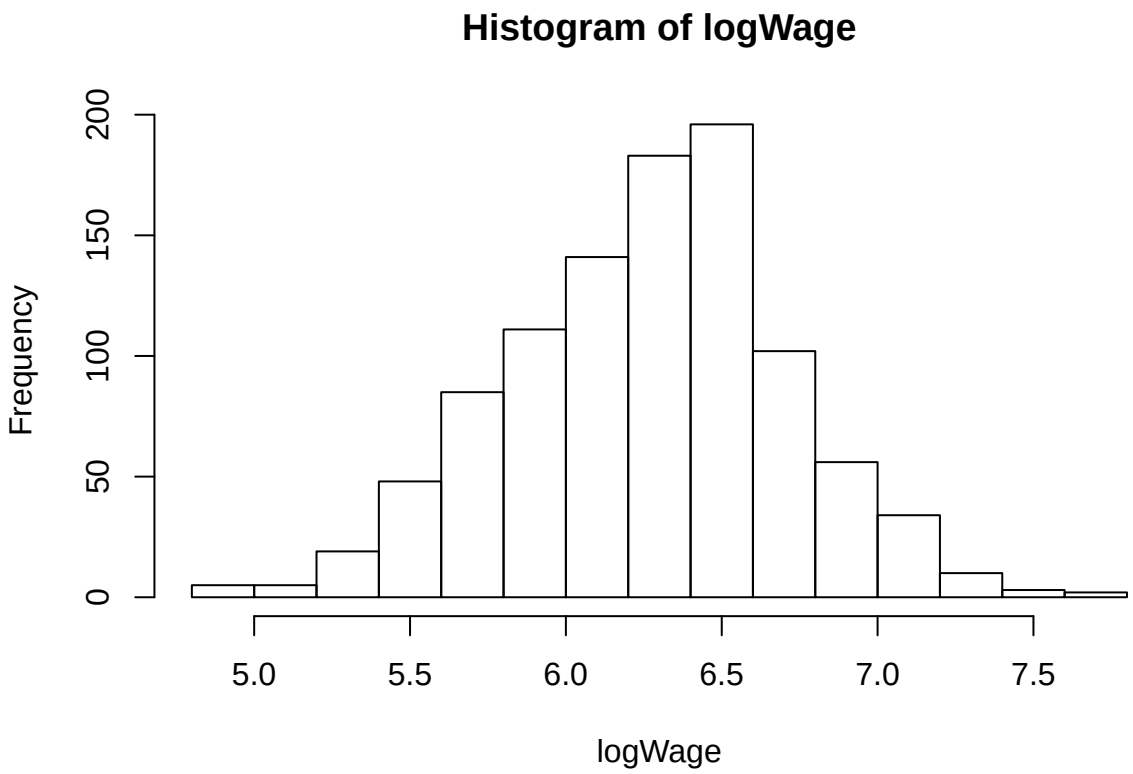
```
summary(logWage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.844   5.991   6.297   6.263   6.555   7.785
```

```
str(logWage)
```

```
##  num [1:1000] 6.86 5.66 6.23 6.47 5.42 ...
```

```
hist(logWage)
```




```
summary(raceColor)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.238   0.000   1.000
```

```
table(raceColor)
```

```
## raceColor
##   0   1
## 762 238
```

City + rural > 1000, so some people identify as both city and rural

```
summary(city)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   0.712   1.000   1.000
```

```
table(city)
```

```
## city
##   0   1
## 288 712
```

```r
summary(rural)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.391   1.000   1.000
```

```r
table(rural)
```

```
## rural
##   0   1
## 609 391
```

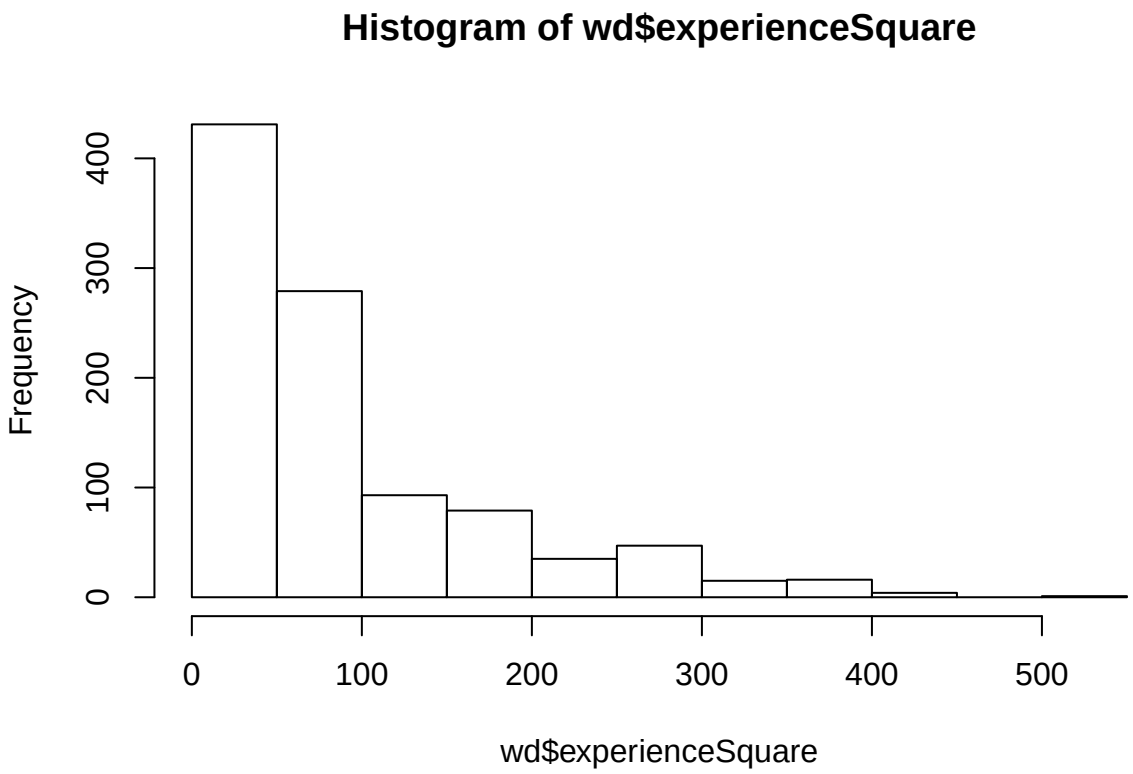
```r
table(z1)
```

```
## z1
##   0   1
## 560 440
```

```r
table(z2)
```

```
## z2
##   0   1
## 314 686
```

```r
wd$experienceSquare = experience**2

hist(wd$experienceSquare)
```




```
attach(wd)
```

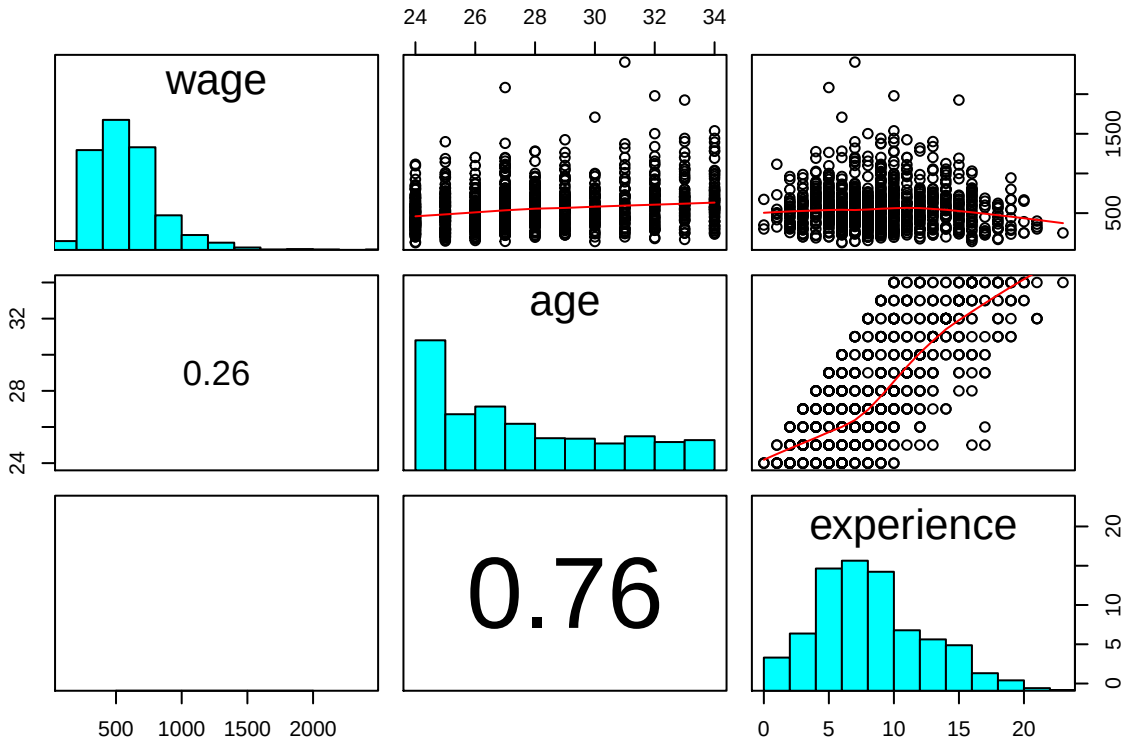
```
## The following objects are masked from wd (pos = 3):
##
##     age, city, dad_education, education, experience, IQscore,
##     logWage, mom_education, raceColor, rural, wage, X, z1, z2
```

Question 4.2

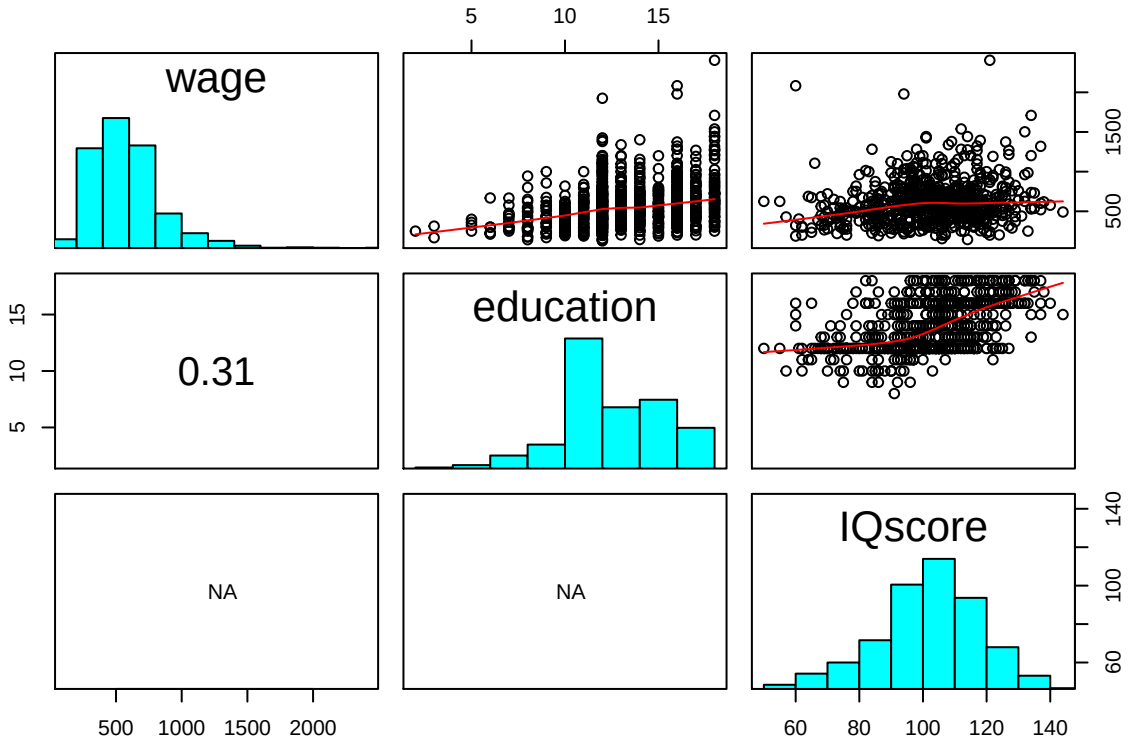
Conduct a bivariate analysis (using tables, graphs, descriptive statistics found in the last 7 lectures) of *wage* and *logWage* and all the other variables in the datasets.

```
panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
```

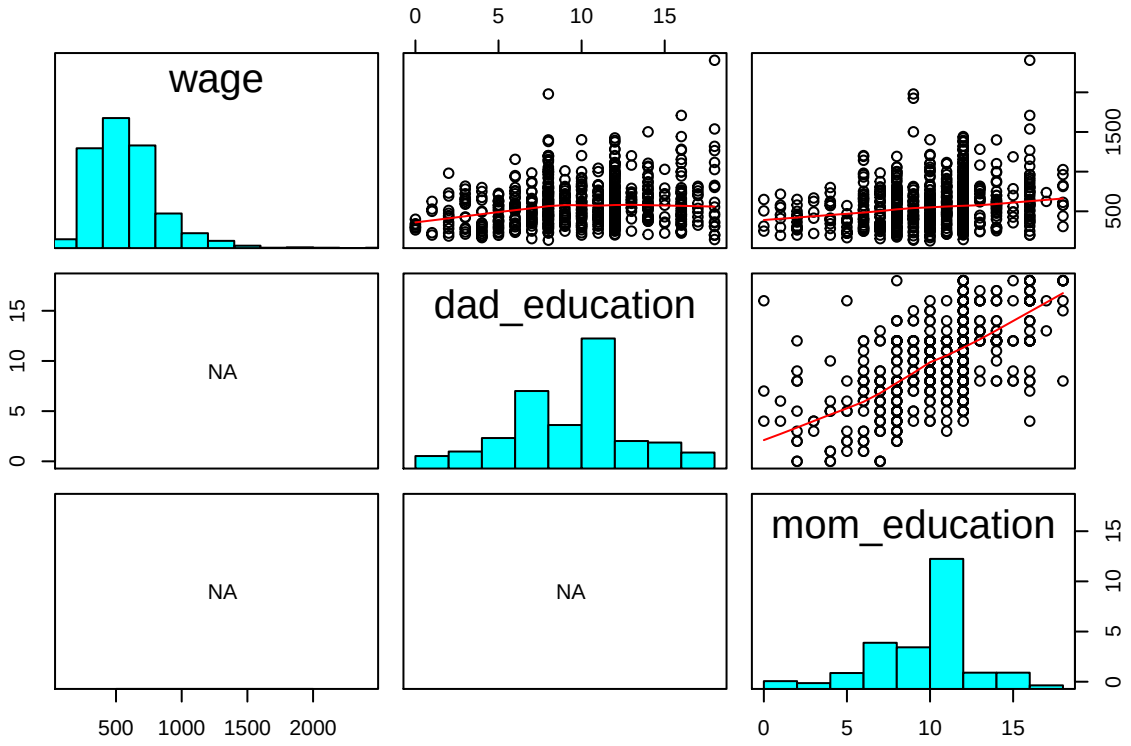
```
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(wage~age+experience,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.panel=panel.his
```



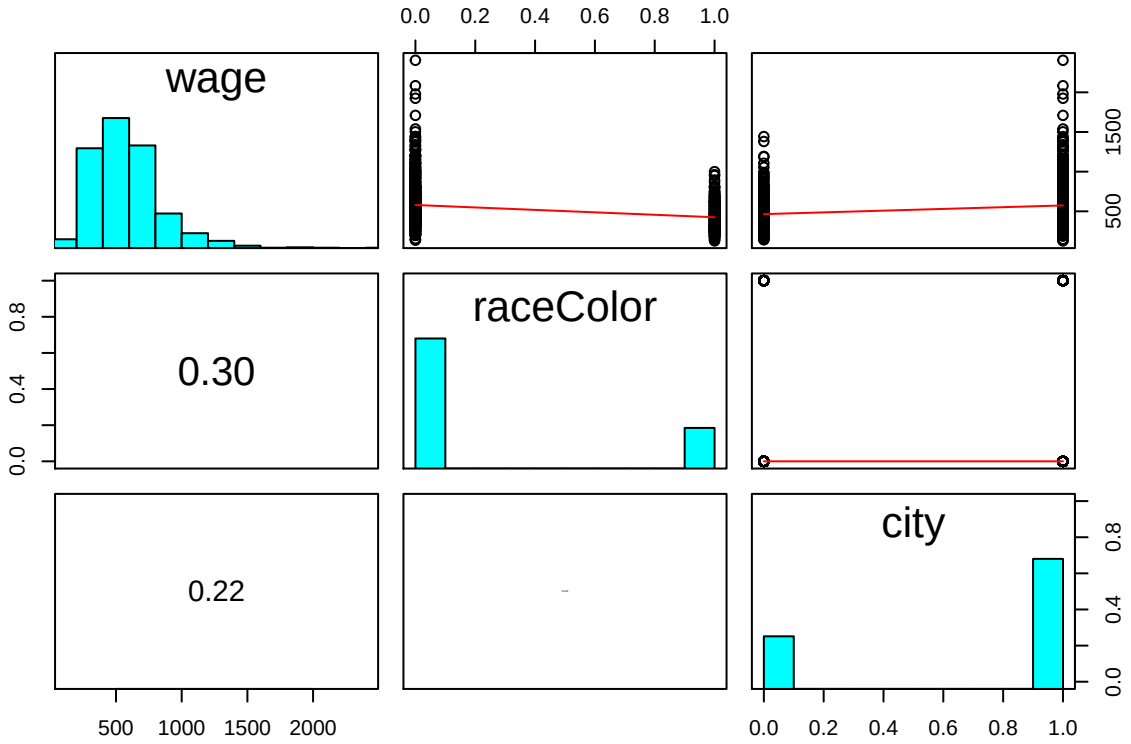

```
pairs(wage~education+IQscore,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.panel=panel
```



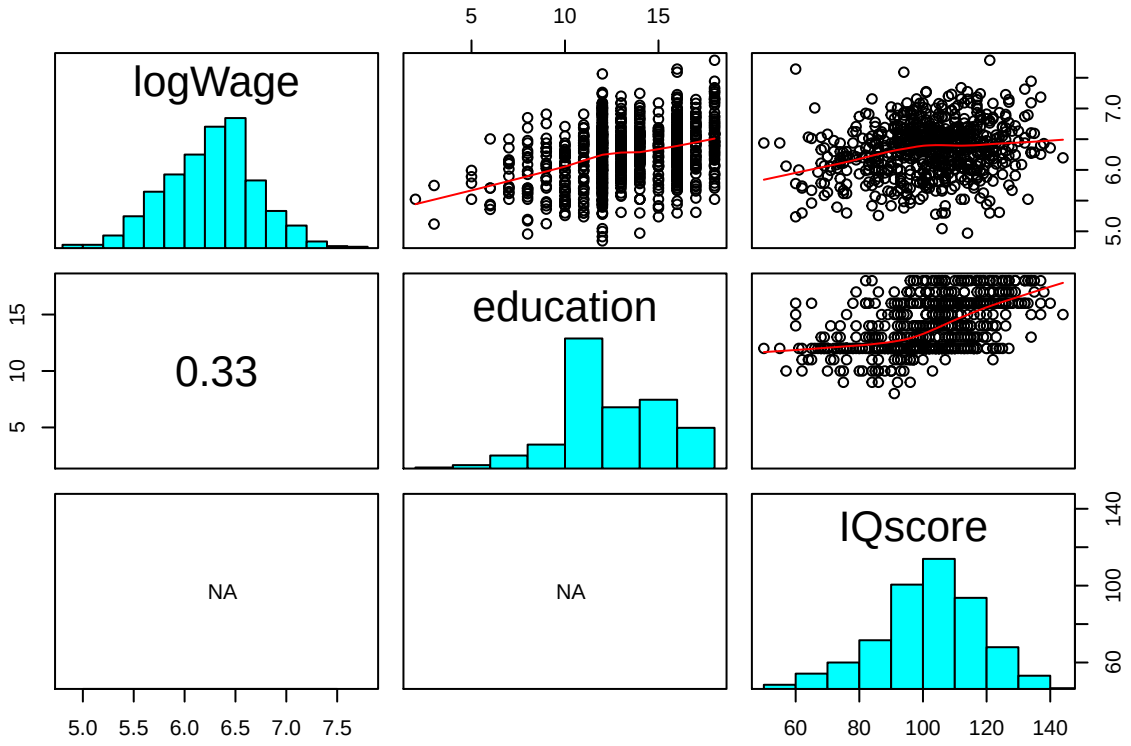

```
pairs(wage~dad_education+mom_education,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.pa
```



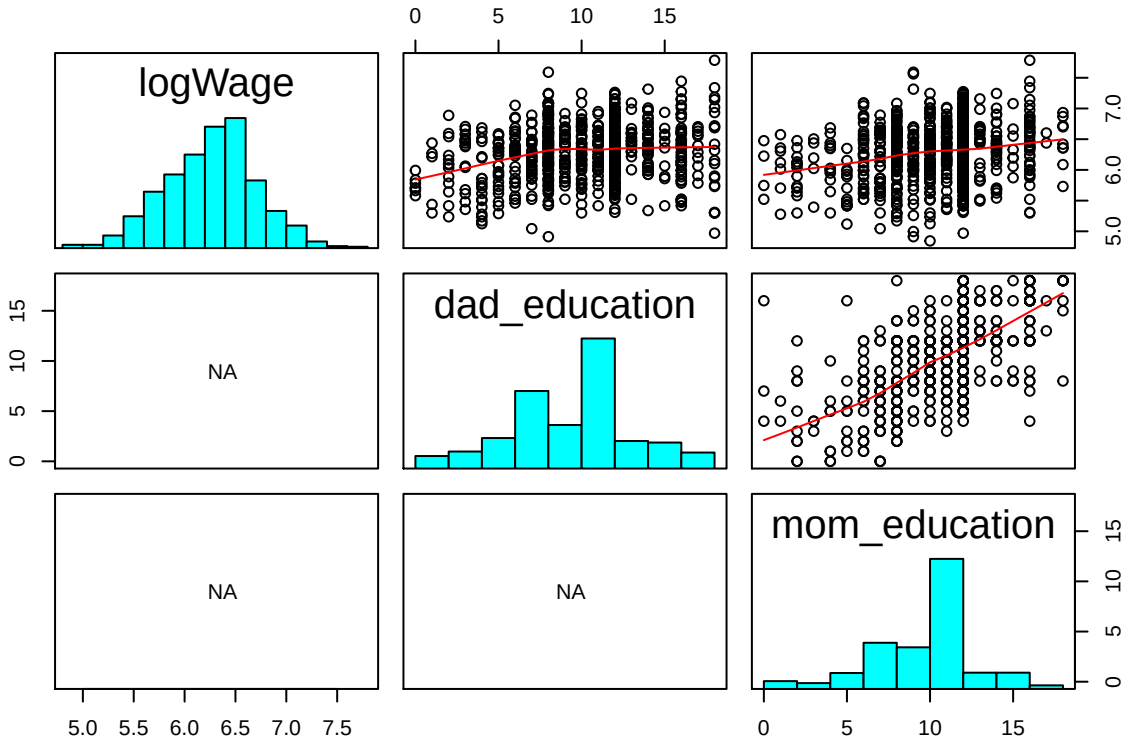

```
pairs(wage~raceColor+city,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.panel=panel.his
```




```
pairs(logWage~education+IQscore,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.panel=pa
```




```
pairs(logWage~dad_education+mom_education,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag
```



Question 4.3

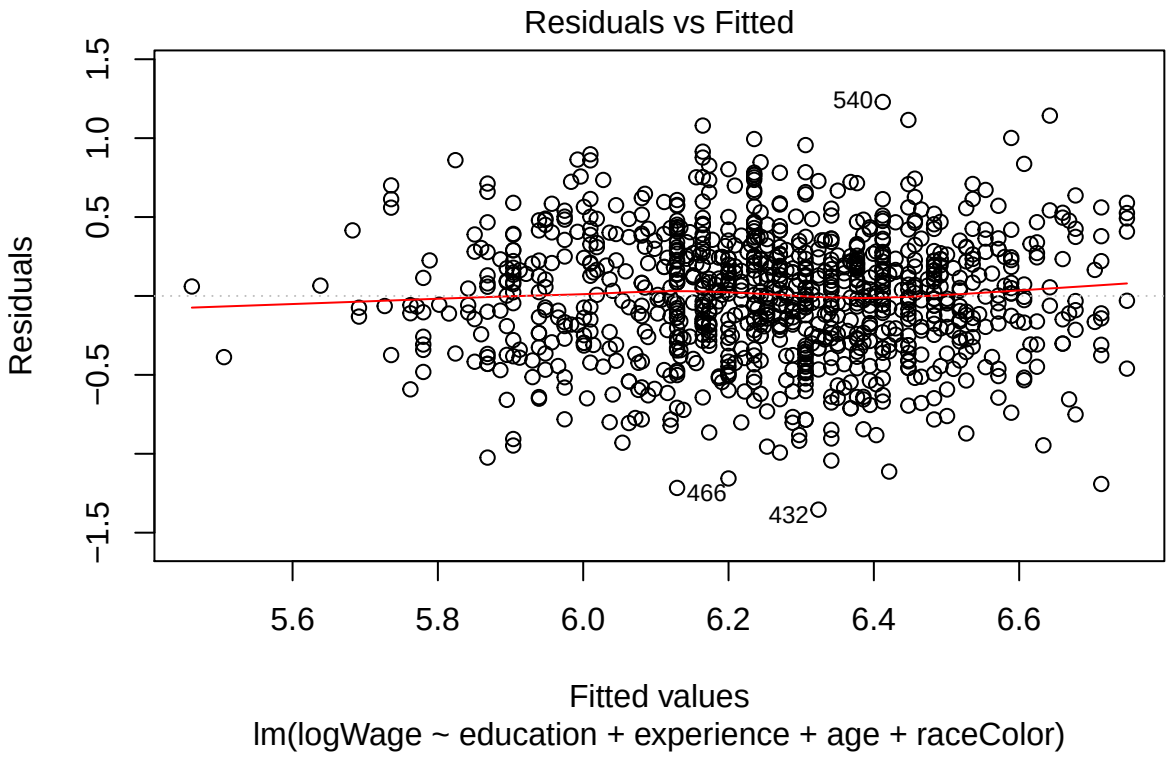
Regress $log(wage)$ on education, experience, age, and raceColor.

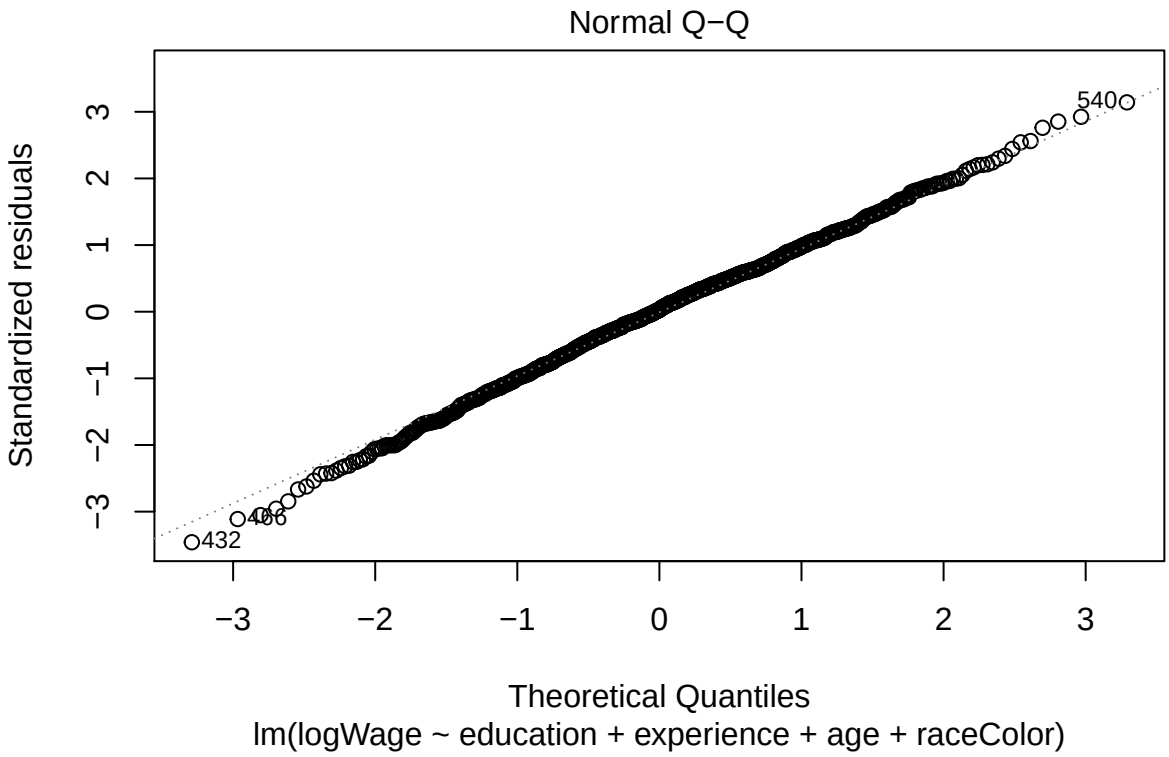
1. Report all the estimated coefficients, their standard errors, t-statistics, F-statistic of the regression, R^2, *adjusted* R^2, and degrees of freedom.

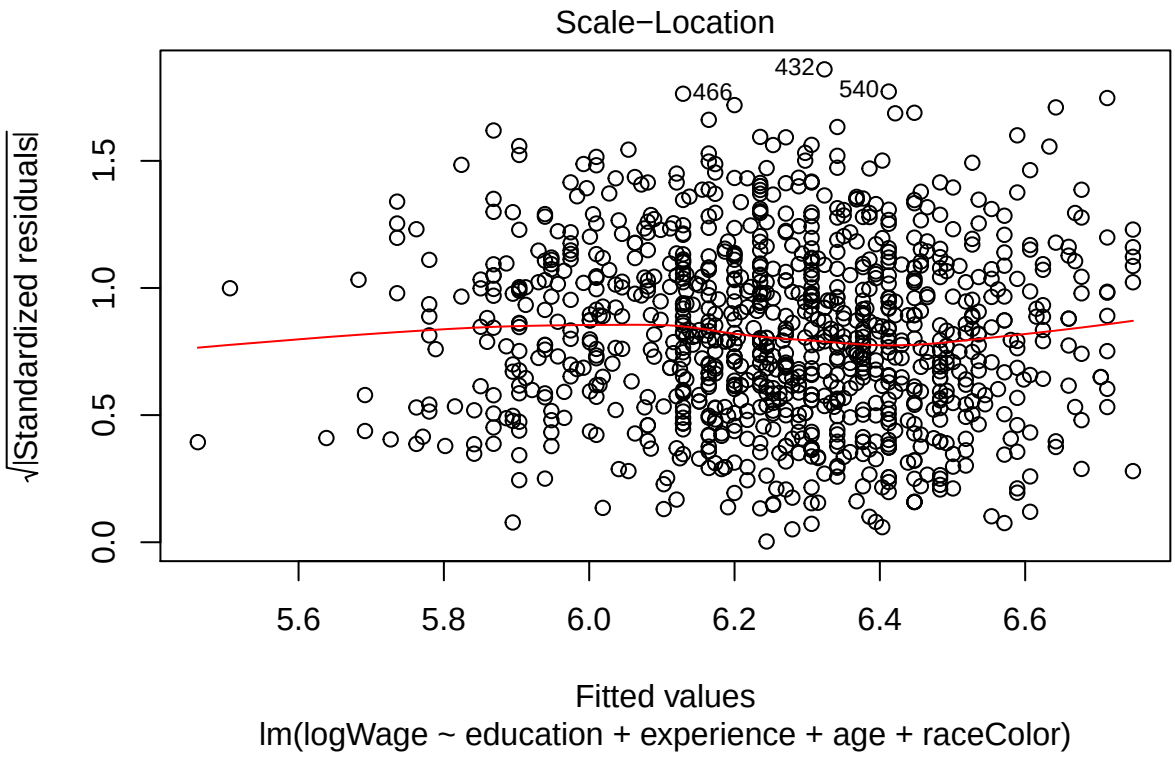
```
model1 = lm(logWage~education+experience+age+raceColor, data=wd)
coeftest(model1)
```

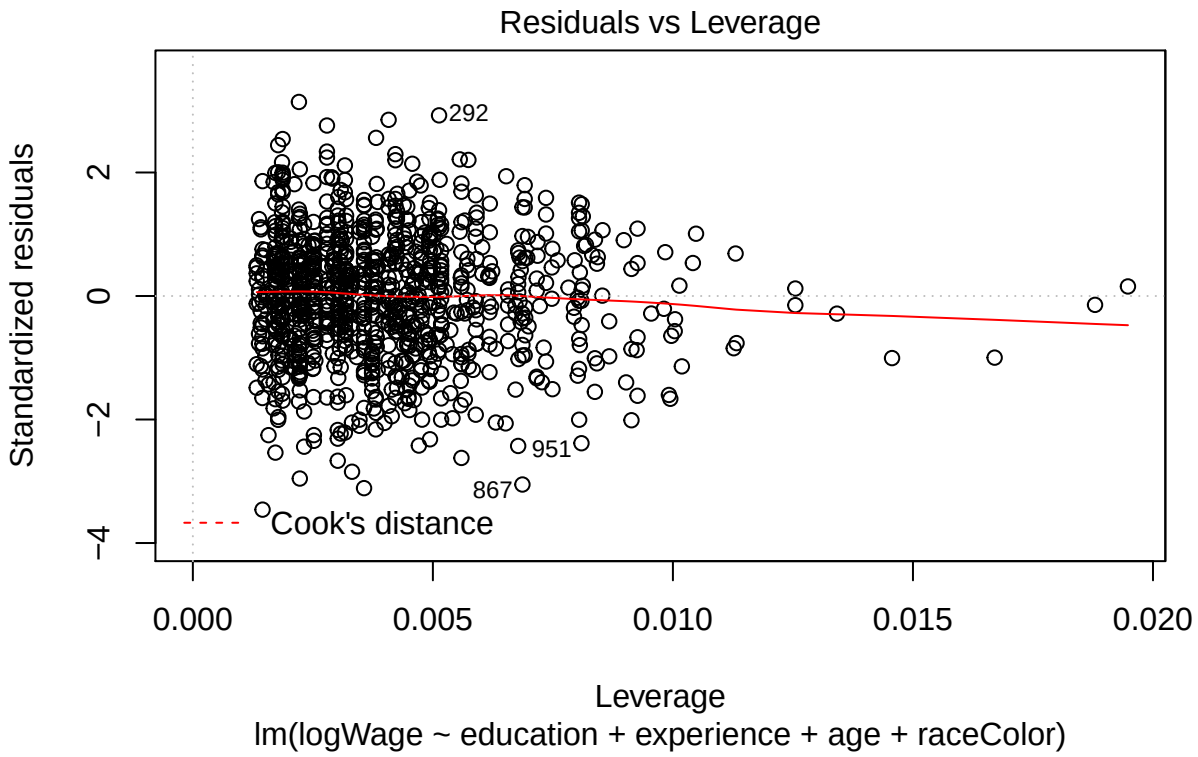
```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  4.9616614  0.1133460 43.7745 < 2.2e-16 ***
## education    0.0796077  0.0063760 12.4856 < 2.2e-16 ***
## experience   0.0353717  0.0039883  8.8689 < 2.2e-16 ***
## raceColor   -0.2608129  0.0304532 -8.5644 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(model1)
```


Residuals vs Fitted
Residuals
1.5
1.0
0.5
−0.5
−1.5
540
466
432
5.6
5.8
6.0
6.2
6.4
6.6
Fitted values
lm(logWage ~ education + experience + age + raceColor)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(logWage ~ education + experience + age + raceColor)
540
432
−3
−2
−1
0
1
2
3


Scale−Location
432
466
540
√|Standardized residuals|
0.0
0.5
1.0
1.5
5.6
5.8
6.0
6.2
6.4
6.6
Fitted values
lm(logWage ~ education + experience + age + raceColor)




```
summary(model1)
```

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

diagnostic plots show homoskedasticity and zero-conditional mean assumptions are satisfied. Errors are

normally distributed, but in a sample size this large this is less important. Residual vs Leverage plot show no points approaching the cook's distance.

Using the summary function to display parameters (not necessary to use the heteroskedasticity-robust versions here)

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

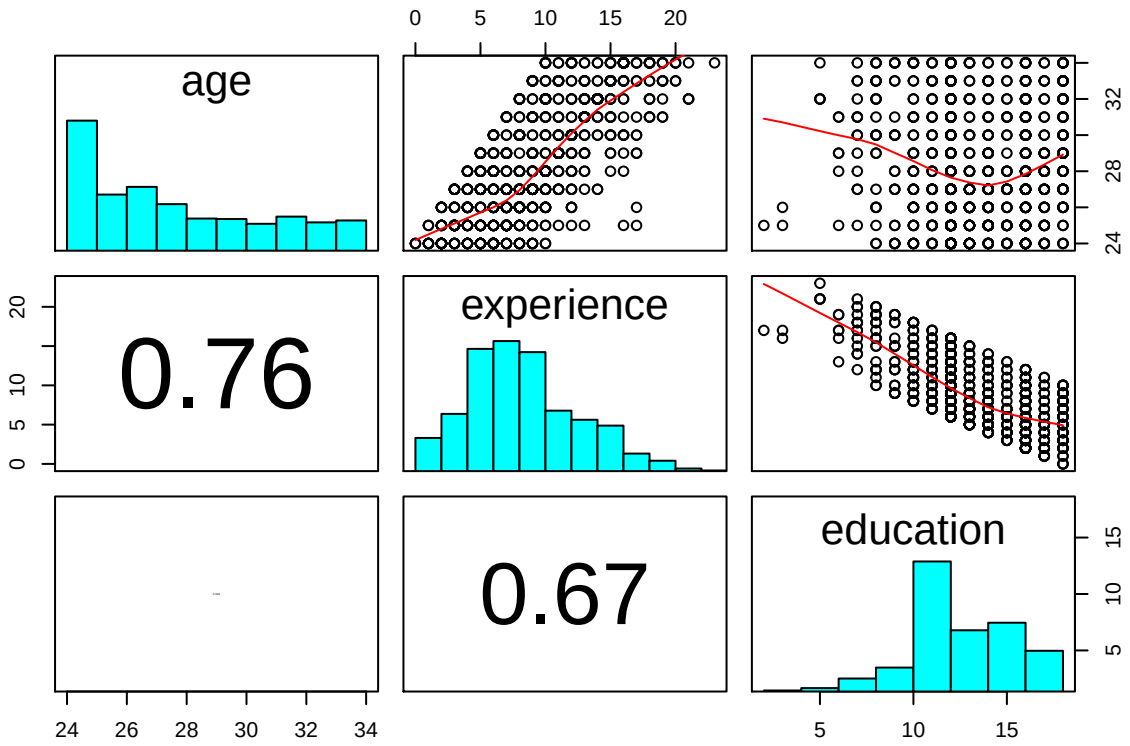
The residual standard error has 996 degrees of freedom which is (n - k -1) n= number of observations k = number of coefficients excluding intercept, in other words we are estimating k+1 parameters

the F-statistic is the ratio of the explained R-squared to the unexplained. The numerator degrees of freedom = # of coeffients being estimated. Denominator df = #of observations - k -1

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).

3 The unexpected result is that R did not calculate an intercept for the age variable. Upon closer examination, this is not surprising. Experience is directly derived from age in this dataset, and the two are highly positively correlated as can be seen from the graph. To correct for this, remove age from the regression model

```
pairs(age~experience+education,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.panel=pan
```




```
model2 = lm(logWage~education+experience+raceColor, data=wd)
summary(model2)
```

```
##
```

```
## Call:
## lm(formula = logWage ~ education + experience + raceColor, data = wd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35396 -0.25550  0.01074  0.24867  1.22932
##
## Coefficients:
##              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 ***
## 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
```

4. Interpret the coefficient estimate associated with education

The coeff on education is ~ 0.08, meaning that an increase in 1 year of education leads to an 8% increase in wages, holding experience and raceColor fixed.

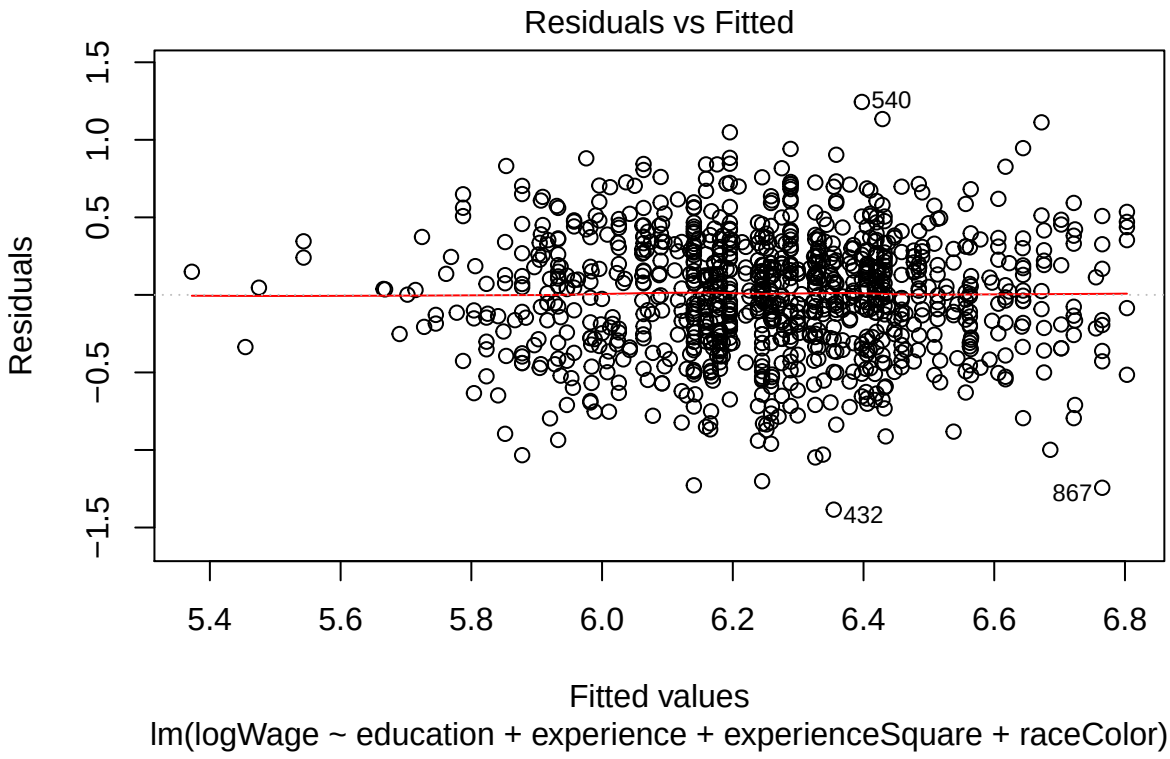
5. Interpret the coefficient estimate associated with experience

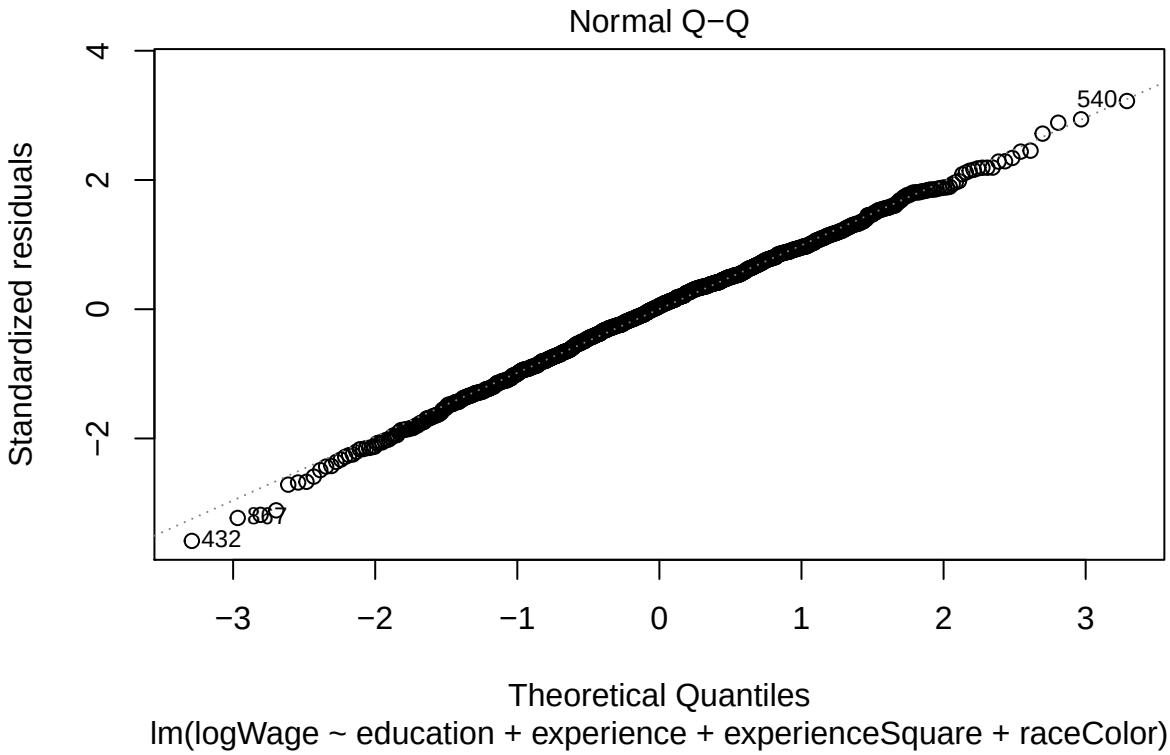
the coeff on experience is 0.03, meaning that an extra year of experience leads to a 3% increase in wages, holding education and raceColor fixed.

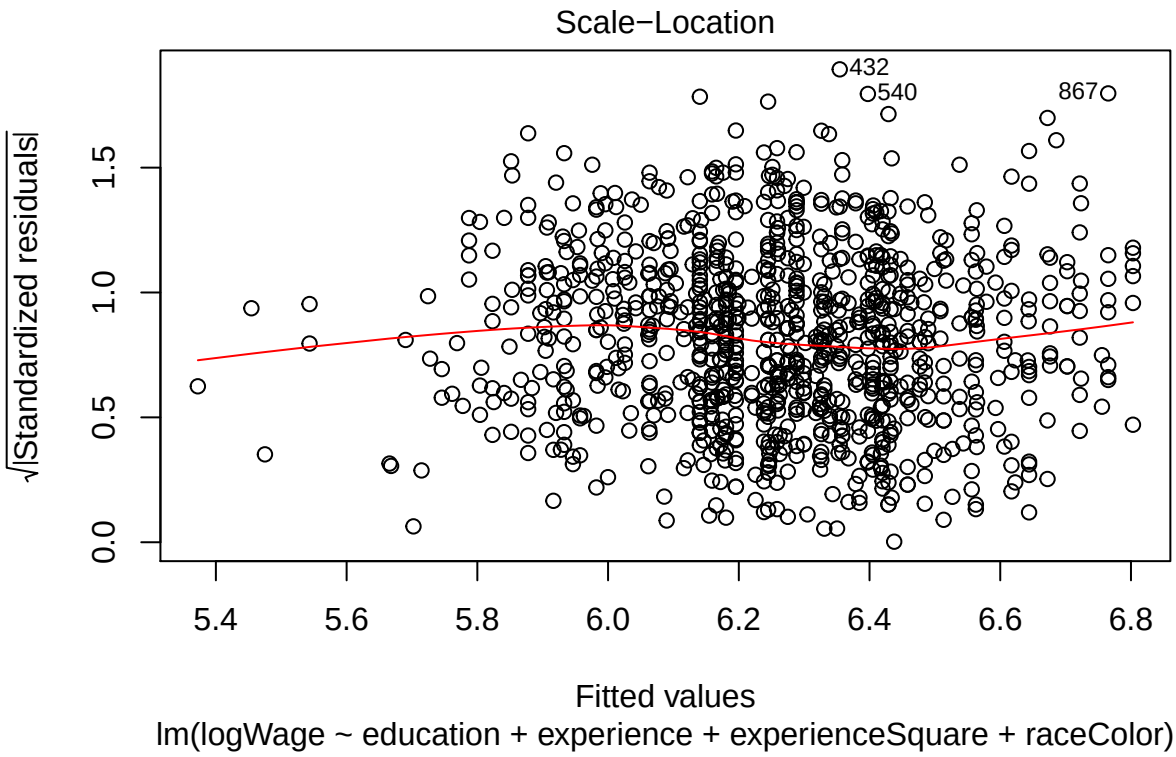
Question 4.4

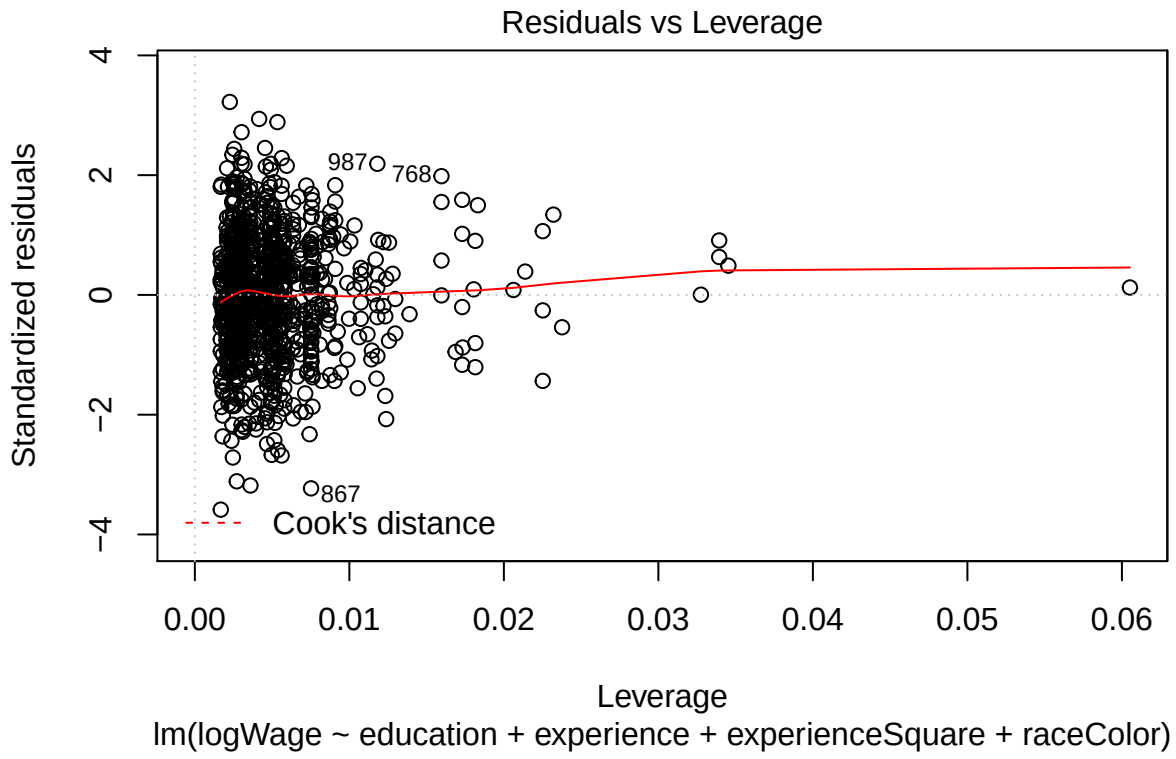
Regress $log(wage)$ on education, experience, experienceSquare, and raceColor.

```
model3 = lm(logWage~education+experience+experienceSquare+raceColor, data=wd)
plot(model3)
```


Residuals vs Fitted
Residuals
1.5
1.0
0.5
−0.5
−1.5
540
867
432
5.4
5.6
5.8
6.0
6.2
6.4
6.6
6.8
Fitted values
lm(logWage ~ education + experience + experienceSquare + raceColor)


Normal Q−Q
Standardized residuals
4
2
0
−2
540
867
432
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(logWage ~ education + experience + experienceSquare + raceColor)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
432
540
867
5.4
5.6
5.8
6.0
6.2
6.4
6.6
6.8
Fitted values
lm(logWage ~ education + experience + experienceSquare + raceColor)




```
coeftest(model3)
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       4.7355175  0.1197719 39.5378 < 2.2e-16 ***
## education         0.0794641  0.0062917 12.6299 < 2.2e-16 ***
## experience        0.0924930  0.0115148  8.0326 2.685e-15 ***
## experienceSquare -0.0028779  0.0005452 -5.2786 1.598e-07 ***
## raceColor        -0.2627226  0.0300528 -8.7420 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model3)
```

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

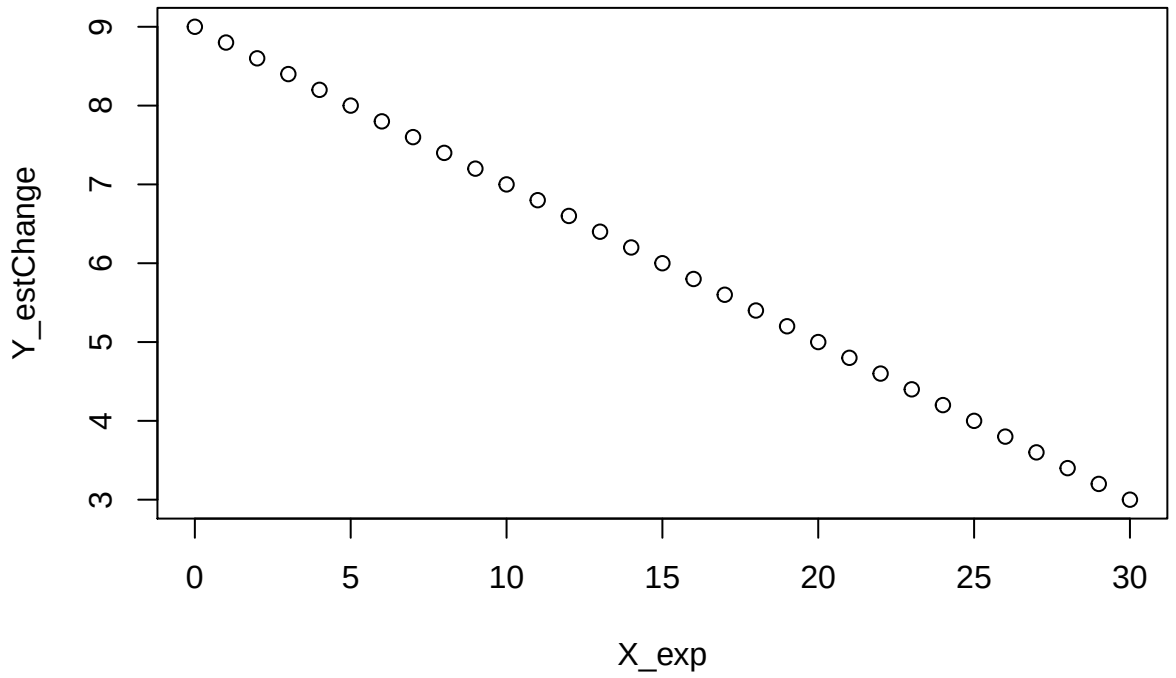
the model is:

logWage = Beta_0 + B_1 * education + B_2 * experience + B_3 * experienceSquare + B_4*raceColor

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

To get the effect of experience on wage, take the partial derivate of the model wrt experience, so we get: d/dE (logWage) = 0.09 -0.002*experience

```
X_exp = seq(0,30)
Y_estChange = (0.09 - X_exp*0.002)*100
plot(X_exp, Y_estChange)
```



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

change in wage when experience=10 yrs: 7% increase (0.09 - 10*0.002)*100

Question 4.5

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

```
model4 = lm(logWage~education+experience+experienceSquare+raceColor+dad_education+mom_education+rural+c
summary(model4)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education + mom_education + rural + city,
##     data = wd)
##
## 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.

4.5.1 from the degrees of freedom on the F-statistic we can see that 714+8+1 = 723 observations out of 1000 were used

```
sum(is.na(wd$dad_education)) # 239
```

```
## [1] 239
```

```r
sum(is.na(wd$mom_education)) # 128
```

```
## [1] 128
```

```r
sum(is.na(wd$mom_education) & is.na(wd$dad_education)): 90
```

```
## [1] 90
```

```r
missing_dad_edc = wd[is.na(wd$dad_education),]

missing_mom_educ = wd[is.na(wd$mom_education),]
```

239+128-90 = 277; 1000 - 277 = 723. This accounts for all the missing observations

could not find any pattern

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

R cannot deal with missing values in a regresion and if we want to find the effect of dad_education and mom_education, we have to throw away the missing values across all variables

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?

```r
wd$dad_educ2 = wd$dad_education
wd$dad_educ2[is.na(wd$dad_educ2)] = mean(wd$dad_education, na.rm=T)
#sum(is.na(wd$dad_educ2))

wd$mom_educ2 = wd$mom_education
wd$mom_educ2[is.na(wd$mom_educ2)] = mean(wd$mom_education, na.rm=T)
#sum(is.na(wd$mom_educ2))

model5 = lm(logWage~education+experience+experienceSquare+raceColor+dad_educ2+mom_educ2+rural+city, data
summary(model5)
```

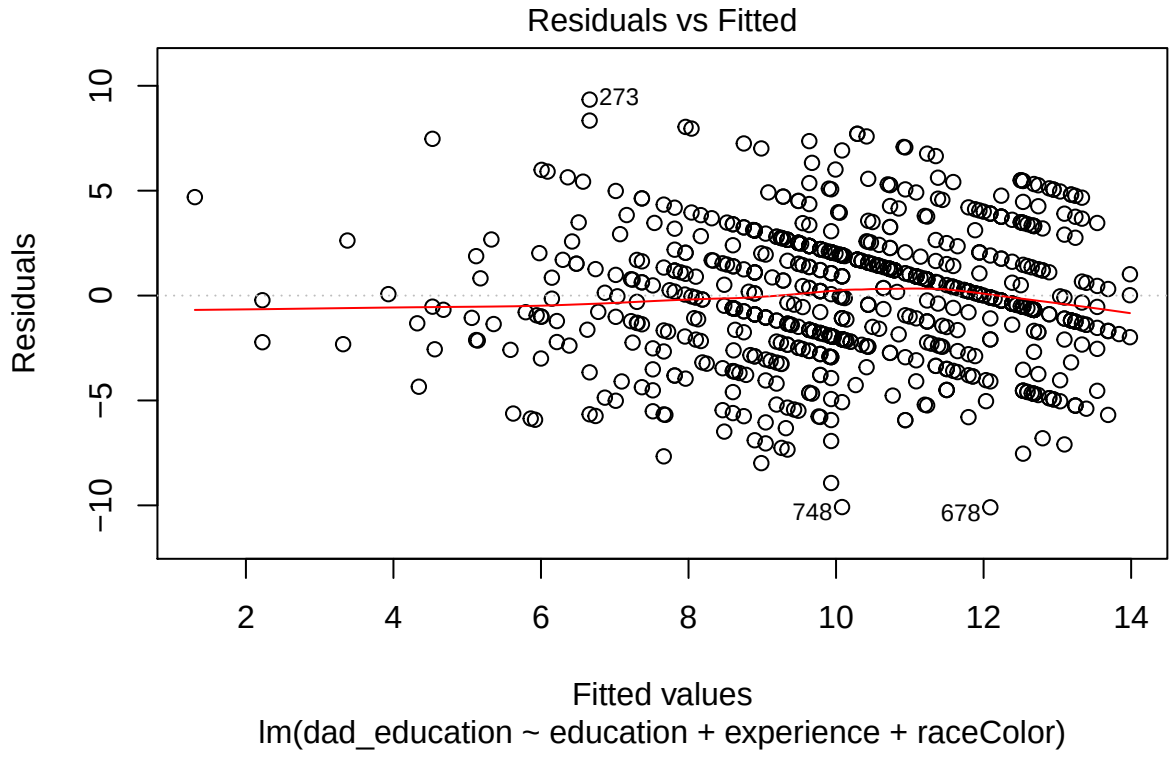
```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_educ2 + mom_educ2 + rural + city, data = wd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30741 -0.23286  0.01943  0.24786  1.28807
##
## Coefficients:
##                    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_educ2        -3.513e-05  4.416e-03  -0.008 0.993656
## mom_educ2         3.485e-03  5.009e-03   0.696 0.486742
```

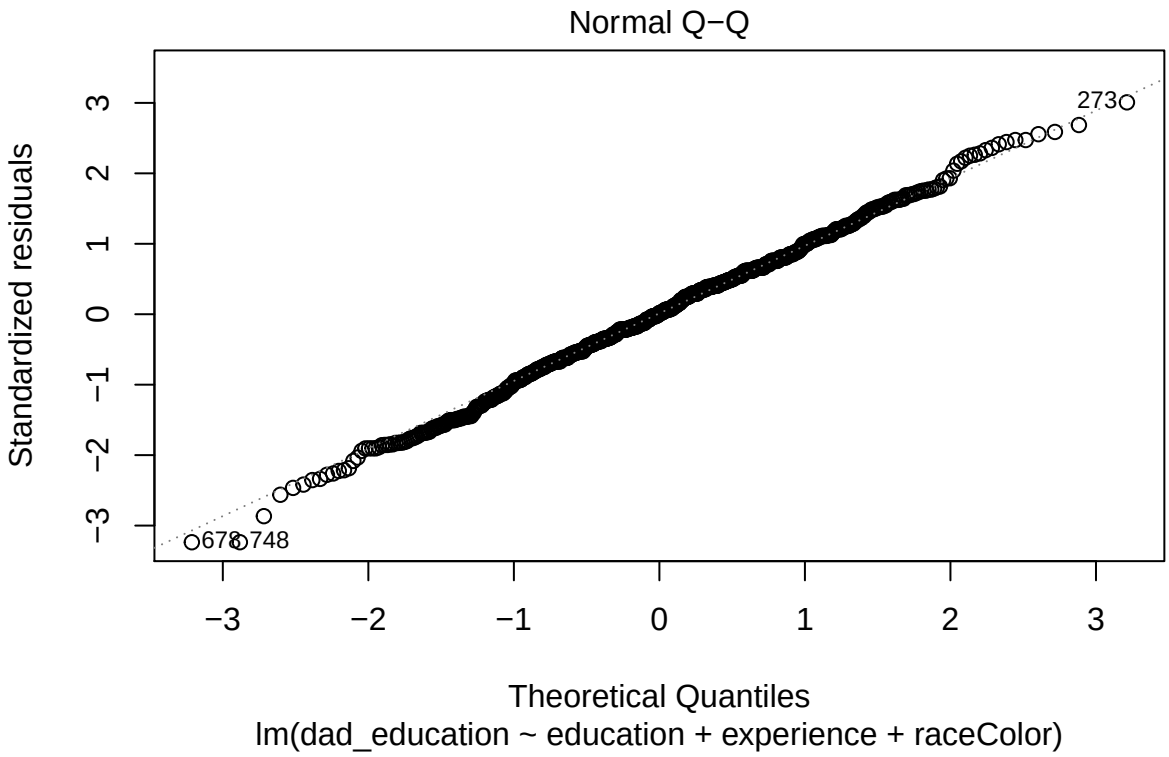
```
## rural            -9.529e-02  2.638e-02  -3.612 0.000319 ***
## city              1.671e-01  2.703e-02   6.183 9.21e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3764 on 991 degrees of freedom
## Multiple R-squared:  0.2981, Adjusted R-squared:  0.2925
## F-statistic: 52.62 on 8 and 991 DF,  p-value: < 2.2e-16
```

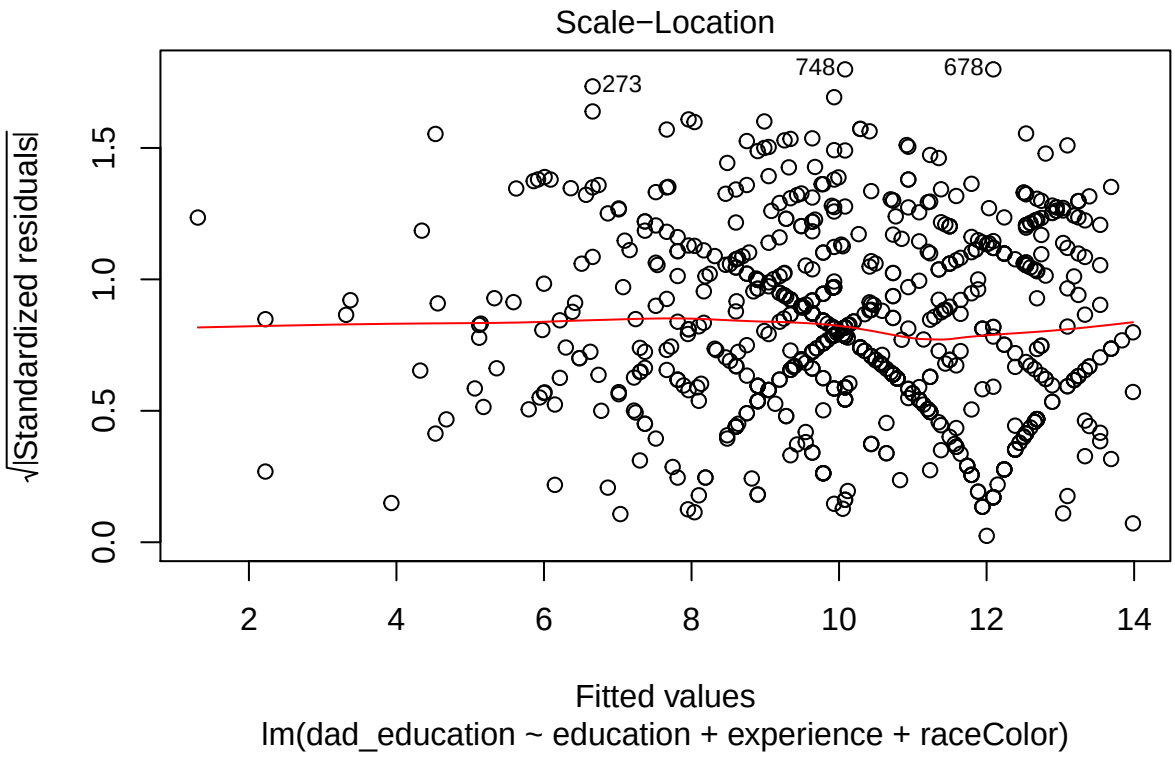
the coefficients on dad_education and mom_education remain statistically insignificant, in fact they dropped in siginificance value

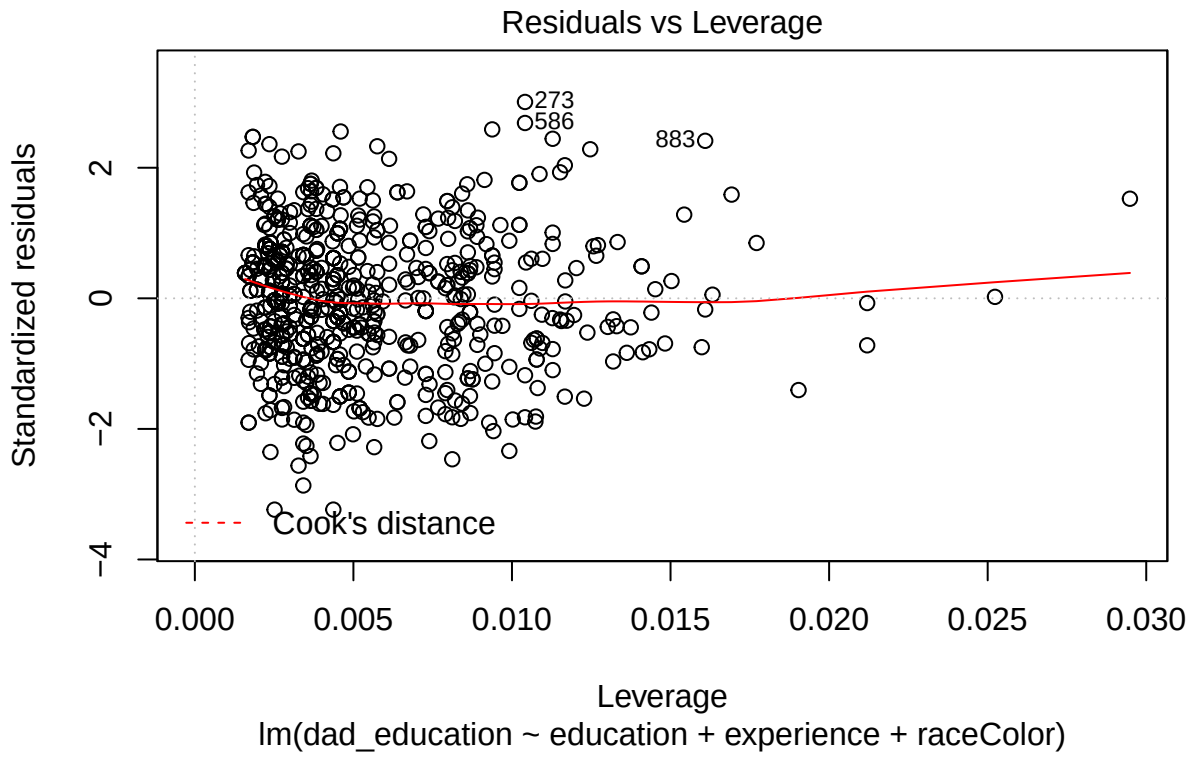
4. How about regress the variable(s) with missing values on education, experience, and raceColor, and use this regression(s) to predict (i.e. "impute") the missing values and then rerun the original regression using all of the observations?

```
model6 =lm(dad_education~education+experience+raceColor, data=wd)
plot(model6)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(dad_education ~ education + experience + raceColor)
273
678
748
−3
−2
−1
0
1
2
3


Scale−Location
273
748
678
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
4
6
8
10
12
14
Fitted values
lm(dad_education ~ education + experience + raceColor)




```
summary(model6)
```

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

dad_educ = 4.93 + 0.5* education -0.148*experience* - *2.12*raceColor

```
wd$dad_educ3 = wd$dad_education
wd_to_fix = wd[is.na(wd$dad_educ3),]
wd_to_fix$dad_educ3 = 4.93 + 0.5 * wd_to_fix$education - 0.148*wd_to_fix$experience - 2.12*wd_to_fix$ra
sum(is.na(wd$dad_educ3))
```

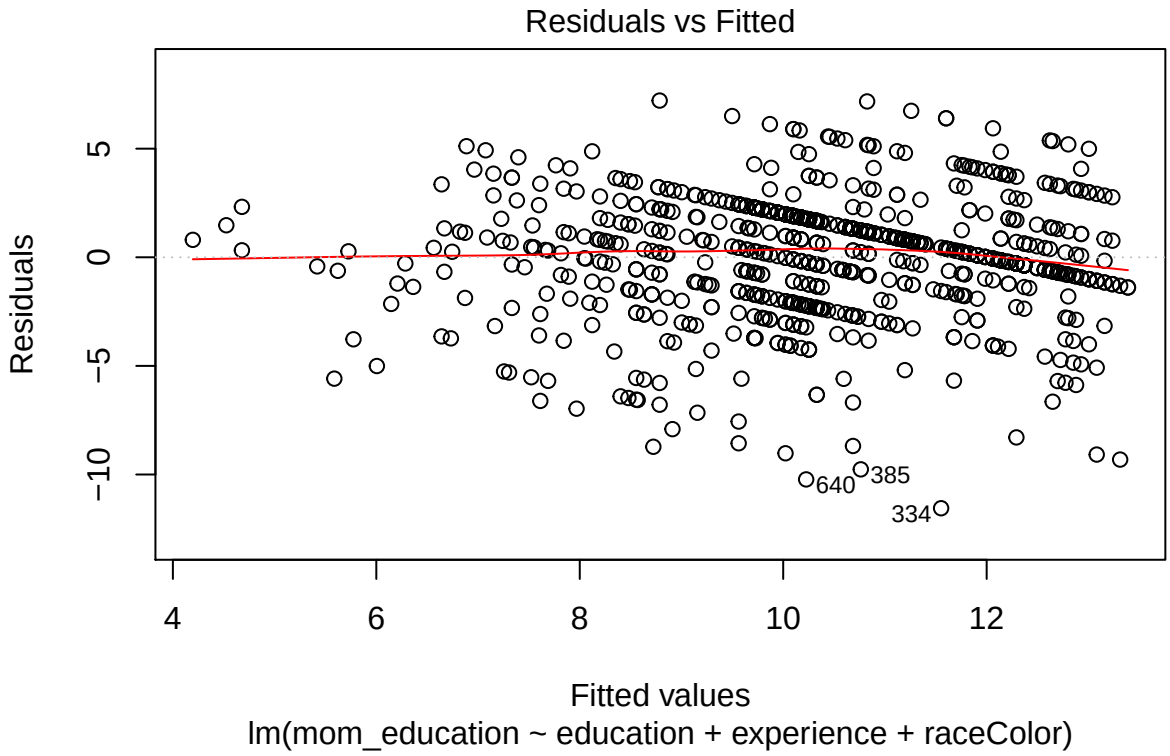
```
## [1] 239
```

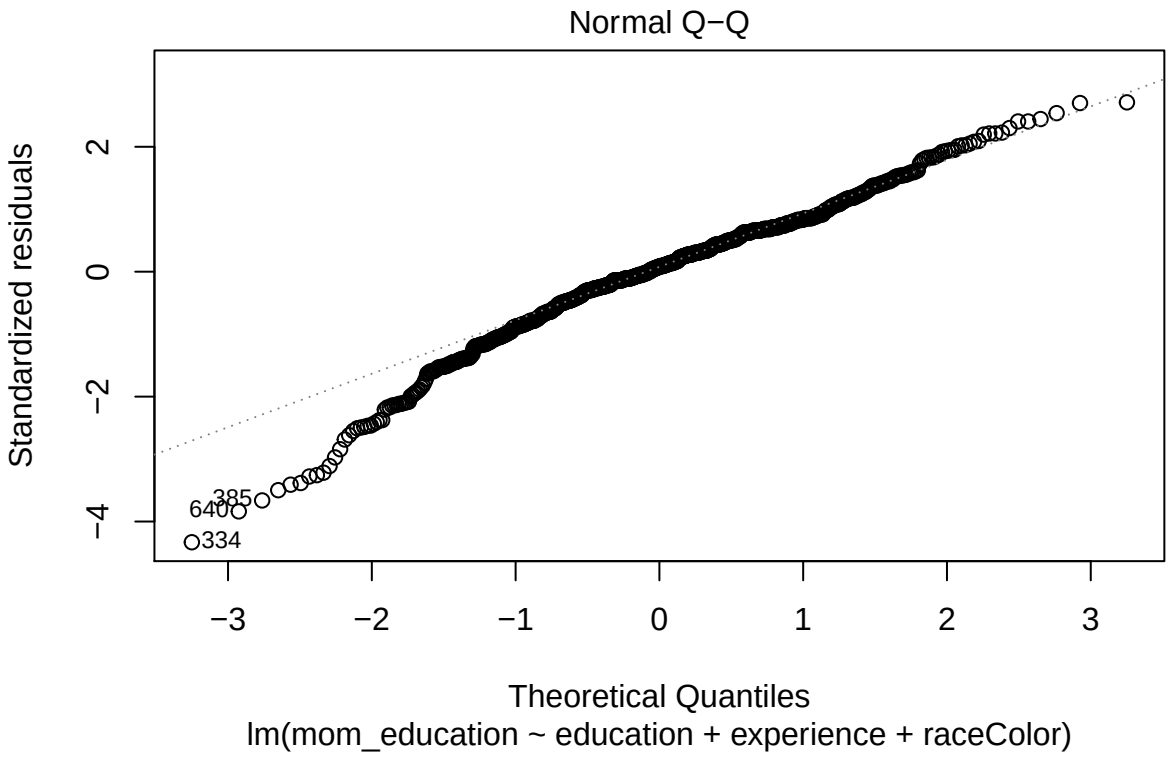
```
sum(is.na(wd_to_fix$dad_educ3))
```

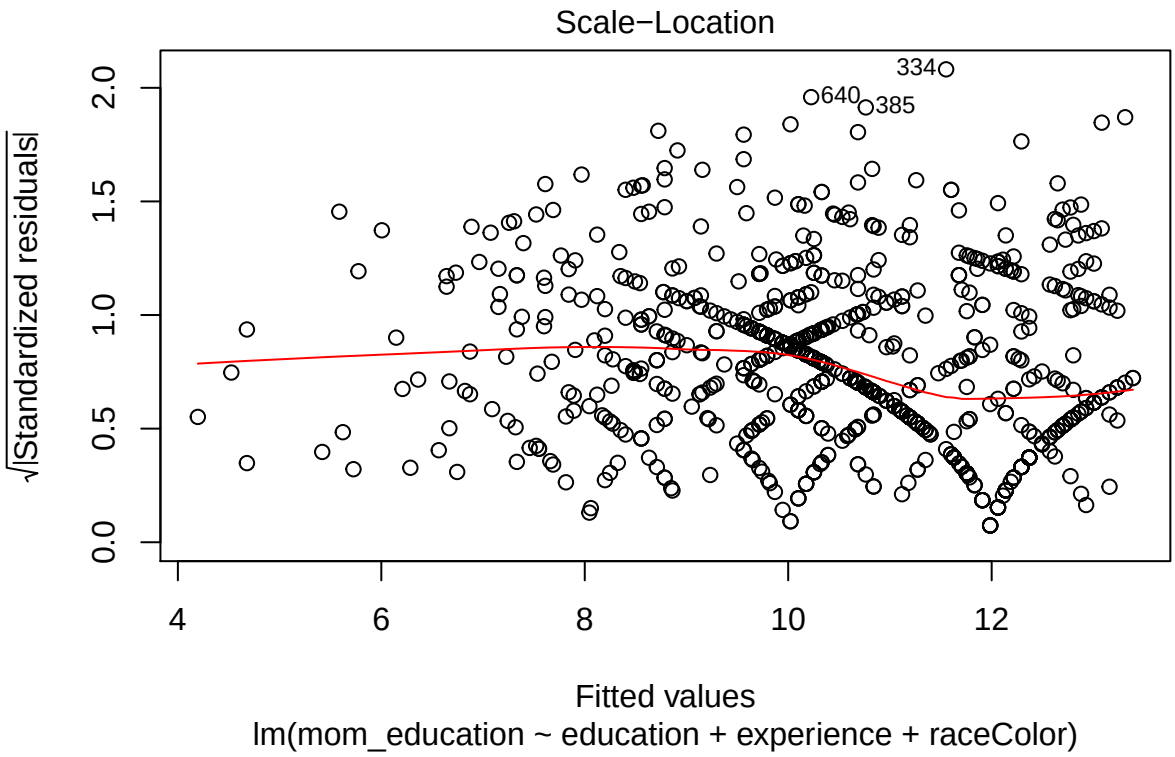
```
## [1] 0
```

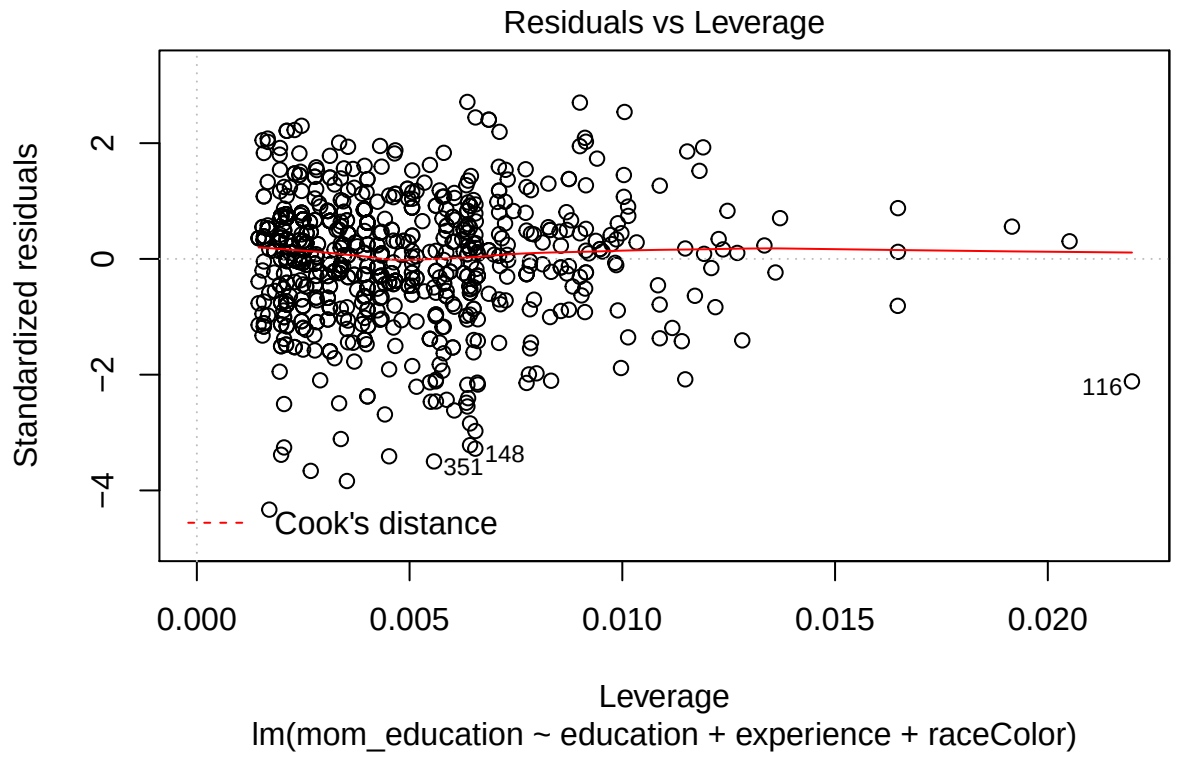
```
wd$dad_educ3[is.na(wd$dad_educ3)] = wd_to_fix$dad_educ3
```

```
model7 =lm(mom_education~education+experience+raceColor, data=wd)
plot(model7)
```




Normal Q−Q
Standardized residuals
2
0
−2
−4
385
640
334
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mom_education ~ education + experience + raceColor)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
334
640
385
4
6
8
10
12
Fitted values
lm(mom_education ~ education + experience + raceColor)




```
summary(model7)
```

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

mom_educ = 5.59 + 0.43* education - 0.07 * experience - 1.46* raceColor

```
wd$mom_educ3 = wd$mom_education
wd_to_fix = wd[is.na(wd$mom_educ3),]
wd_to_fix$mom_educ3 = 5.59 + 0.43*wd_to_fix$education - 0.07*wd_to_fix$experience - 1.46*wd_to_fix$race
sum(is.na(wd$mom_educ3))
```

```
## [1] 128
```

```
sum(is.na(wd_to_fix$mom_educ3))
```

```
## [1] 0
```

```
wd$mom_educ3[is.na(wd$mom_educ3)] = wd_to_fix$mom_educ3
```

```
model8 = lm(logWage~education+experience+experienceSquare+raceColor+dad_educ3+mom_educ3+rural+city, data
summary(model8)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_educ3 + mom_educ3 + rural + city, data = wd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30591 -0.22956  0.01781  0.24775  1.28275
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.726851   0.121274  38.977  < 2e-16 ***
## education         0.070086   0.006689  10.479  < 2e-16 ***
## experience        0.089490   0.011223   7.974 4.22e-15 ***
## experienceSquare -0.002655   0.000532  -4.991 7.10e-07 ***
## raceColor        -0.226505   0.032073  -7.062 3.08e-12 ***
## dad_educ3         0.002375   0.004741   0.501 0.616454
## mom_educ3         0.002381   0.005171   0.460 0.645323
## rural            -0.094837   0.026396  -3.593 0.000343 ***
## city              0.166531   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
```

still not statitically significant effect. The coefficient is 0.2% increase in wage for every extra year of dad or mom education, which is a pretty small effect.

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

4.5.6 Prefer which one? The first one. Truest to data.

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)?

Z1 must be uncorrelated with the error term u

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?

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?

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?

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?

```
library(car)
library(lmtest)

setwd("C:/Subha/WS271-Regression/Labs/lab2_w271_2016Spring")

W = read.csv("Wealthy_candidates.csv")
str(W)
```

```
## '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 ...
```

```
summary(W$urb)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02835 0.08387 0.14660 0.18730 0.24320 0.80230
```

```
summary(W$lit)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2418  0.3846  0.4602  0.4512  0.5105  0.6524
```

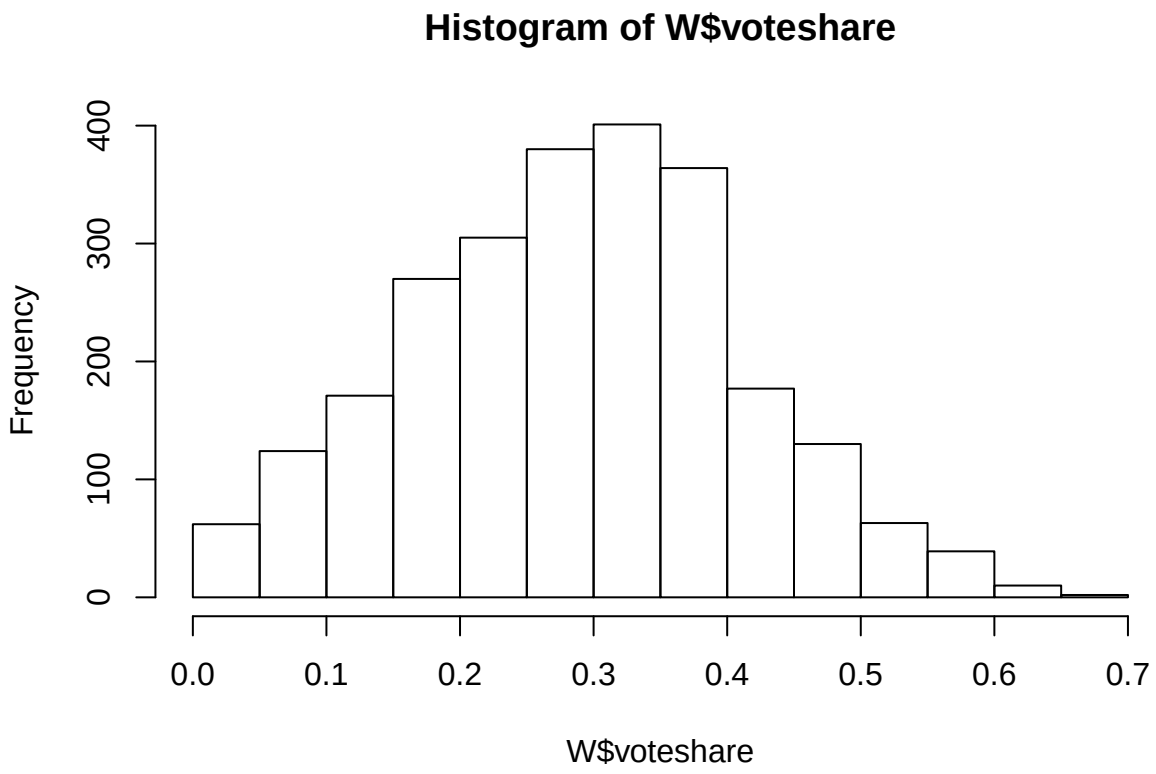
```
summary(W$voteshare)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.006037 0.199600 0.293400 0.287900 0.368000 0.693300
```

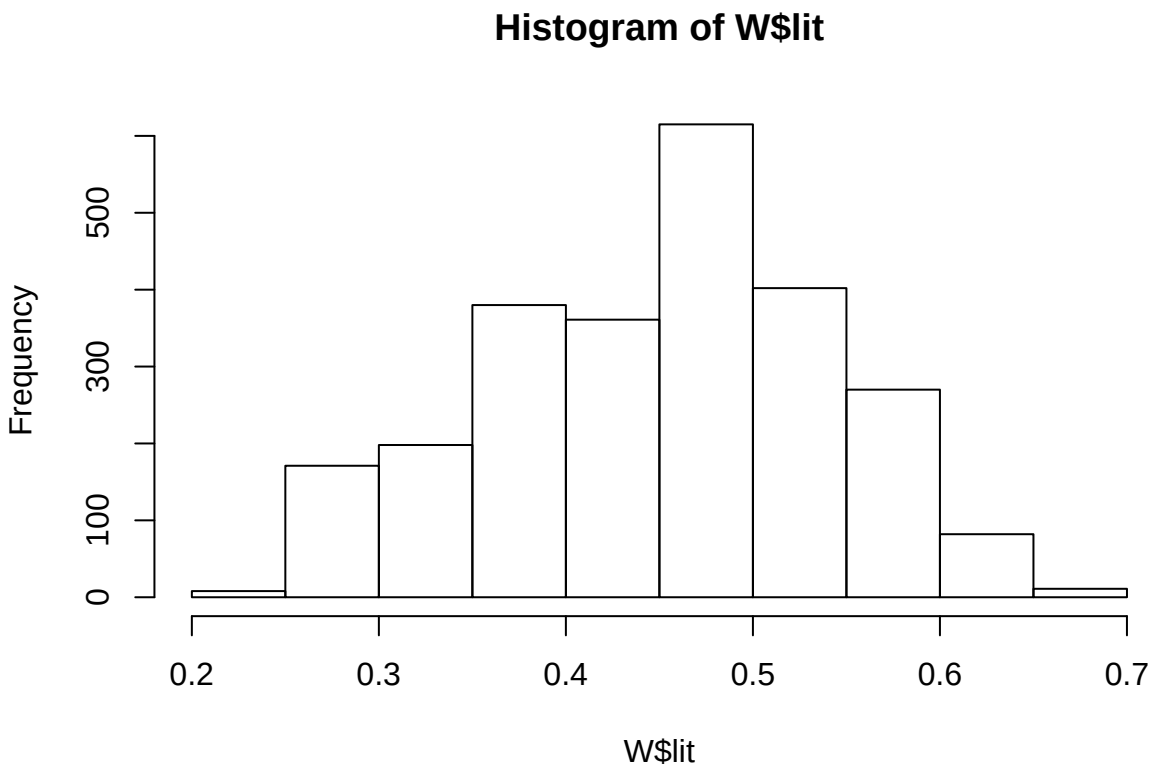
```
summary(W$absolute_wealth)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 2.000e+00 1.875e+05 1.337e+06 5.034e+06 4.092e+06 1.216e+09         1
```

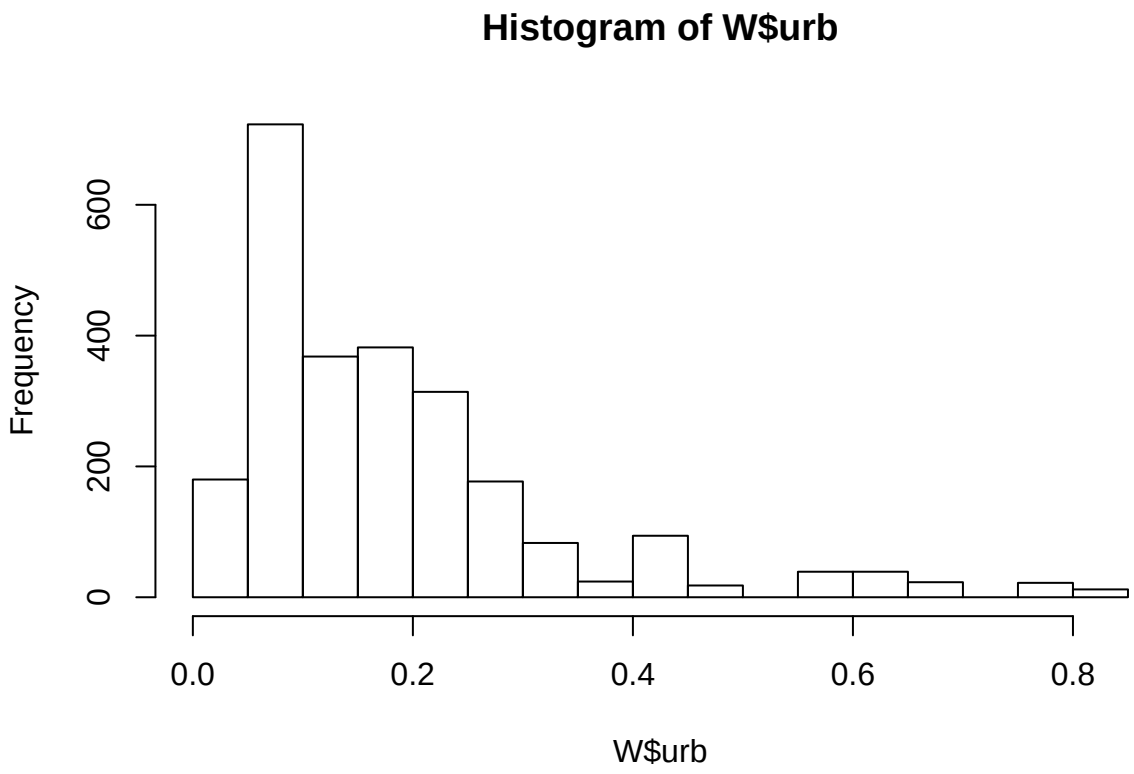
```
hist(W$voteshare)
```



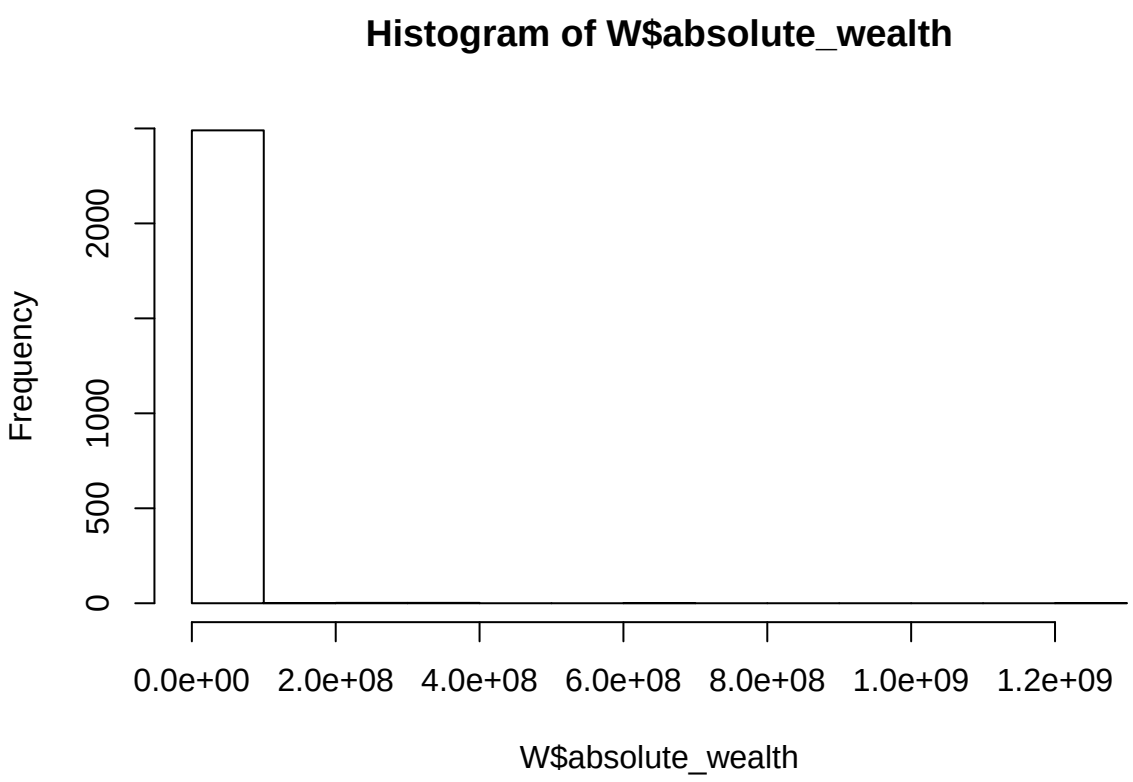

```
hist(W$lit)
```


Histogram of W$lit
Frequency
0
100
300
500
0.2
0.3
0.4
0.5
0.6
0.7
W$lit

```
hist(W$urb)
```




```
hist(W$absolute_wealth)
```




```
sum(W$absolute_wealth > 2e+8, na.rm=T) # 6 values
```

```
## [1] 6
```

```
sum(W$absolute_wealth > 1e+8, na.rm=T) # 7 values
```

```
## [1] 7
```

```
sum(W$absolute_wealth > 5e+7, na.rm=T) #19 values
```

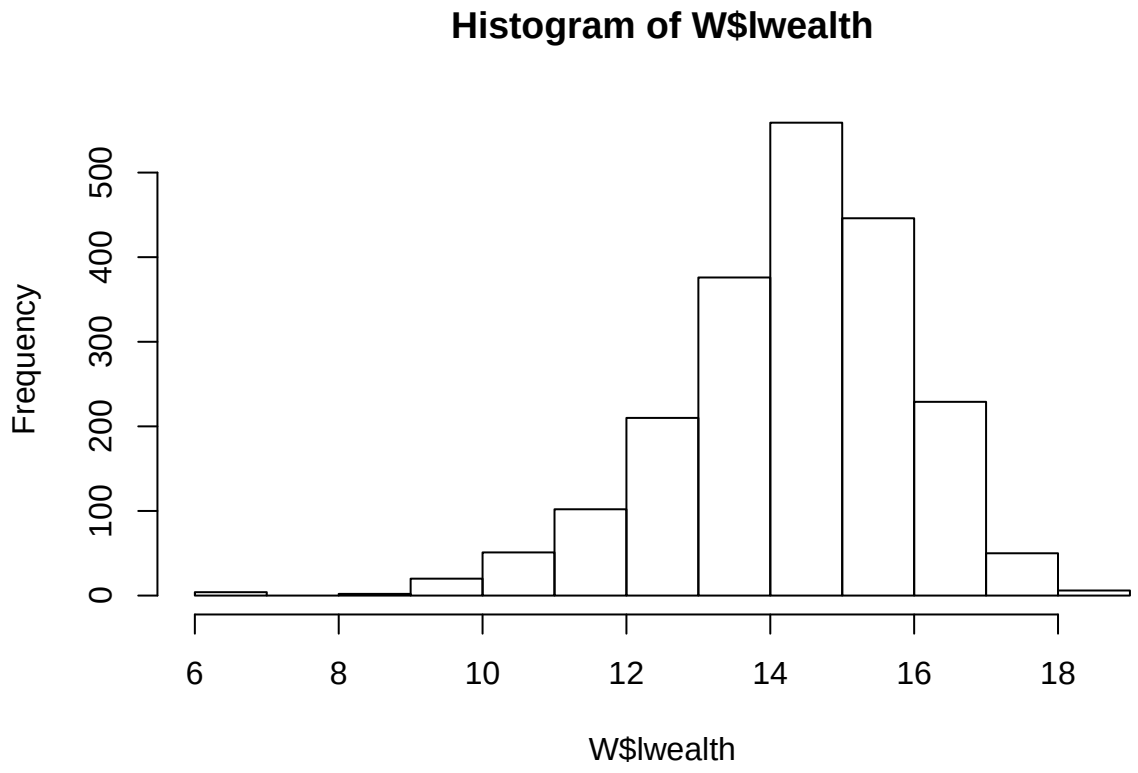
```
## [1] 19
```

```
W[W$absolute_wealth >= 1e+8, ]
```

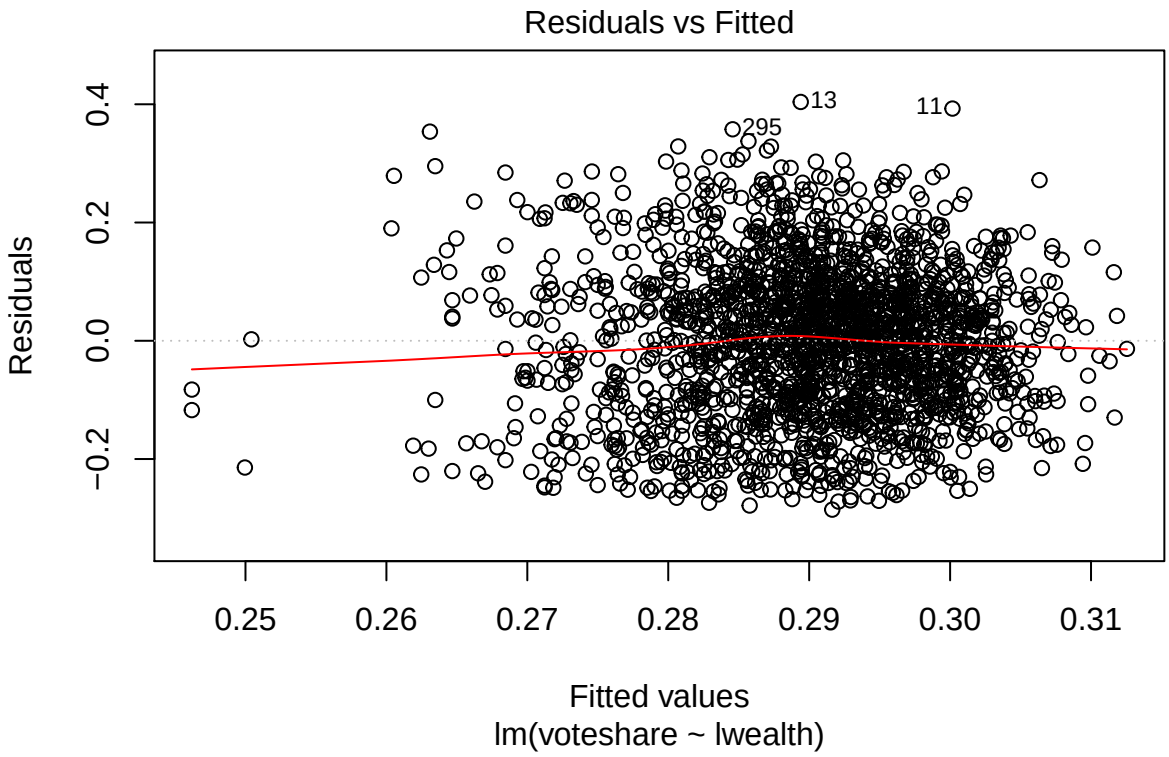
```
##         X   region        urb       lit voteshare absolute_wealth
## NA     NA     <NA>         NA        NA        NA              NA
## 432   432 Region 2 0.06139975 0.2731756 0.1718122       163372416
## 1322 1322 Region 1 0.24590805 0.5078786 0.1901637       699396480
## 1618 1618 Region 1 0.13551243 0.4419104 0.2468476       301821632
## 2089 2089 Region 1 0.07092855 0.3432697 0.4521993       268619840
## 2094 2094 Region 1 0.07391634 0.4784835 0.3156603       209518016
## 2119 2119 Region 1 0.25507668 0.4906439 0.2755198      1216399232
## 2387 2387 Region 1 0.09270675 0.3826344 0.3107426       308832992
```

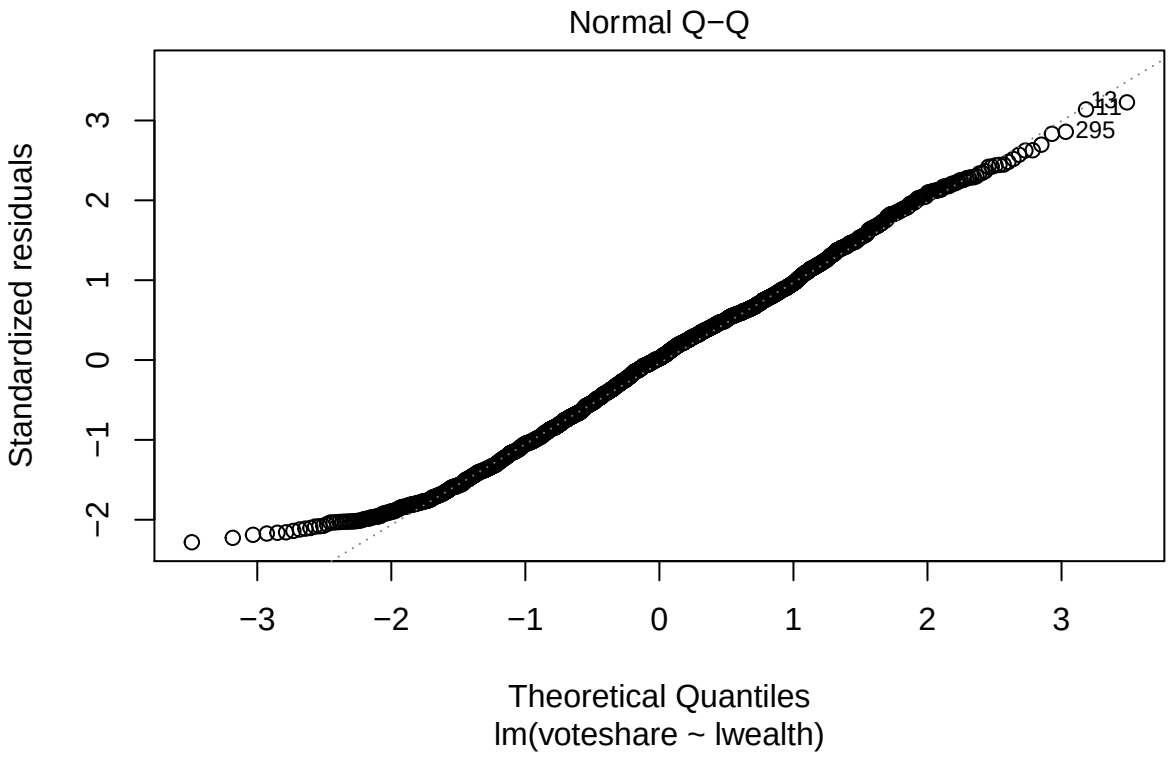
wealth has an expected positive skew. There seem to be very few values beyond 2E+8, and many values at 2. create a new variable with the outliers and missing values taken out. Use the log of the wealth for analysis

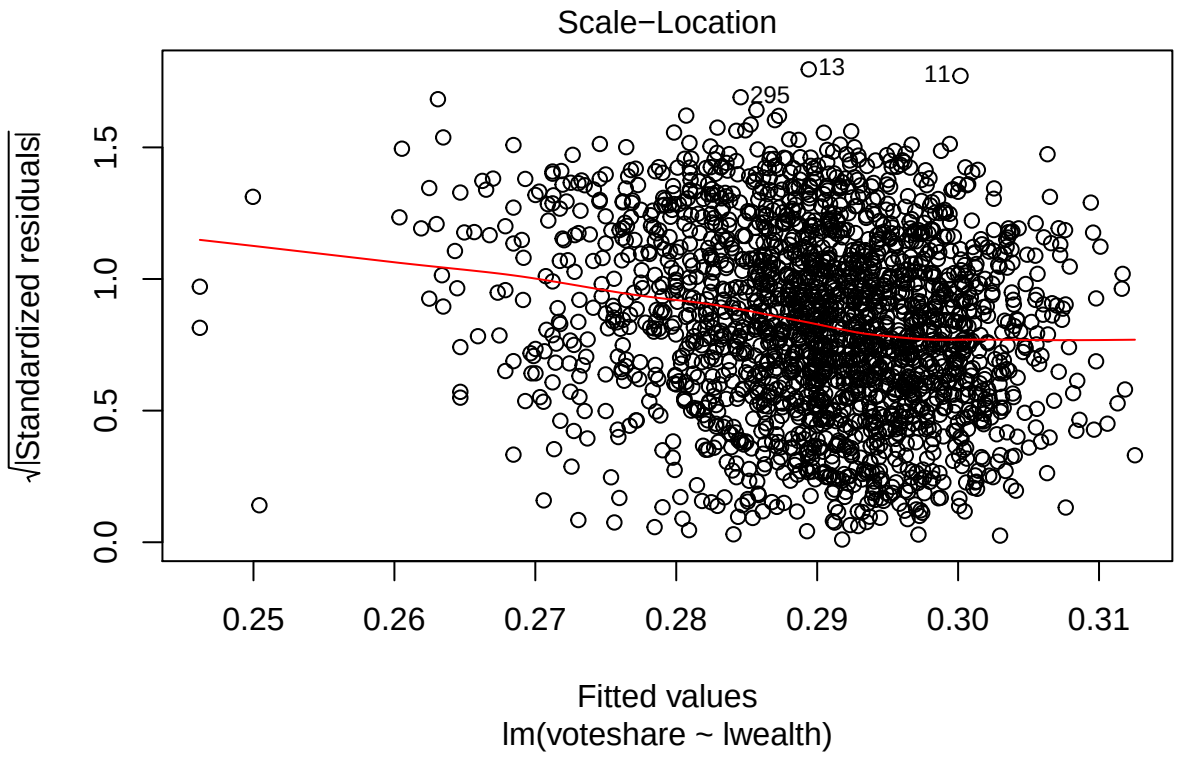
```
W$abs_wealth2 = W$absolute_wealth
W$abs_wealth2[W$abs_wealth2 == 2] = NA
W$abs_wealth2[W$abs_wealth2 > 1E+8] = NA


W$lwealth = log(W$abs_wealth2)
hist(W$lwealth)
```

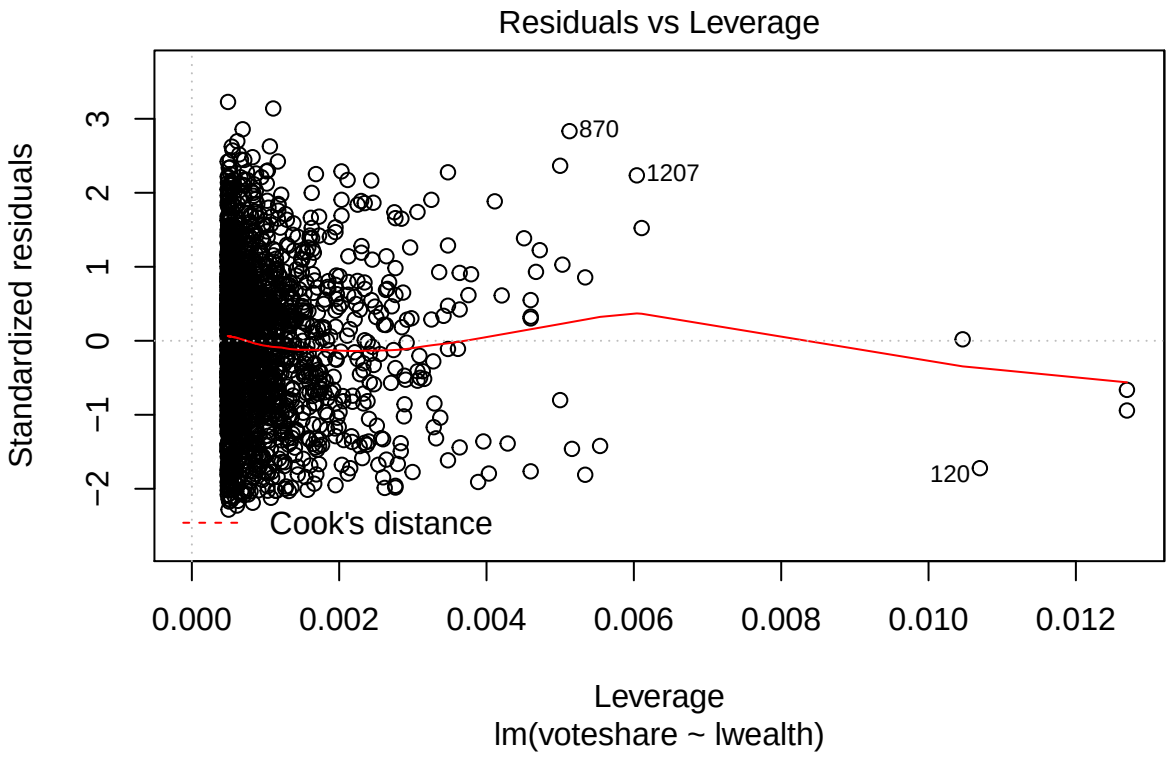



```
model1 = lm(voteshare ~ lwealth, data=W)
plot(model1)
```


Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
13
11
295
0.25
0.26
0.27
0.28
0.29
0.30
0.31
Fitted values
lm(voteshare ~ lwealth)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(voteshare ~ lwealth)
13
11
295
−3
−2
−1
0
1
2
3


Scale−Location
13
11
295
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.25
0.26
0.27
0.28
0.29
0.30
0.31
Fitted values
lm(voteshare ~ lwealth)




```
summary(model1)
```

```
##
## Call:
## lm(formula = voteshare ~ lwealth, data = W)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28560 -0.09061  0.00267  0.08019  0.40392
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.212391   0.024599   8.634  < 2e-16 ***
## lwealth     0.005438   0.001707   3.185  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1252 on 2053 degrees of freedom
##   (443 observations deleted due to missingness)
## Multiple R-squared:  0.004918,   Adjusted R-squared:  0.004434
## F-statistic: 10.15 on 1 and 2053 DF,  p-value: 0.001467
```

The model is statistically significant. the coefficient on logWealth is 0.005, indicating that a 1% increase in wealth increases the voteshare by (0.005/100) %

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?

Adding a quadratic term might help, since we expect diminshing returns from wealth in predicting voteshare.

```
W$lwealth_squared = W$lwealth**2


model2 = lm(voteshare ~ lwealth+lwealth_squared, data=W)
summary(model2)
```

```
##
## Call:
## lm(formula = voteshare ~ lwealth + lwealth_squared, data = W)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28834 -0.09009  0.00229  0.08012  0.40060
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.0125225  0.1244502  -0.101   0.9199
## lwealth          0.0385926  0.0180647   2.136   0.0328 *
## lwealth_squared -0.0012031  0.0006526  -1.844   0.0654 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1251 on 2052 degrees of freedom
##   (443 observations deleted due to missingness)
## Multiple R-squared:  0.006564,   Adjusted R-squared:  0.005595
## F-statistic: 6.779 on 2 and 2052 DF,  p-value: 0.001163
```

With the addition of the quadratic, the coefficient on lWealth has reduced, and the quadratic term has a negative coefficient indicating diminishing returns. Wealthier candidates do fare better, but only upto the turning point, which is (B_1/2*B_2) =

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.

Add region as a factor. Create dummy variables for them

```
W$region2 = W$region == "Region 2"
summary(W$region2)
W$region3 = W$region == "Region 3"
summary(W$region3)

model3 = lm(voteshare~lwealth+region2+region3, data=W)
#plot(model3)
summary(model3)

anova(model1, model3)
```

We find a statistically significant result for the effect of region and wealth on voteshare. The adjusted R-squared is quite low at 3.8%.

To compare the 2 models, use anova. We see that the second model is significantly different from the first one. The addition of the region variables is a relevant addition to the model.

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?

Model1 found a statistically significant effect of wealth on voteshare, but this does not imply the model is causal or unbiased.

In order to be causal, the lWealth variable (i.e the candidate's wealth) must be uncorrleated with the error term. This means that no unobserved variables may be correlated with it. We know this to not be true, since many other factors such as family background, business connections, education, etc are correlated with a person's wealth.

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?

voteshare = B_0 + B_1 * lWealth + B_2 * region3 + B_3 * region3 + error

The difference in difference equation would be: delta(voteshare) = delta(B_1) * lWealth + delta(B_2) * region2 + delta(B_3) * region3 + delta(error)

This would yield a causal result if the unobserved variables that are correlated with the predictors are constant at the two time periods when the measurements were taken.

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 sup- ported 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.

```
library(car)
library(lmtest)

setwd("C:/Subha/WS271-Regression/Labs/lab2_w271_2016Spring")


load("RetailSales.Rdata")
rs = retailSales
str(rs)
```

```
## 'data.frame':    84672 obs. of  14 variables:
##  $ Year             : int  2004 2004 2004 2004 2004 2004 2004 2004 2004 2004 ...
##  $ Product.line     : Factor w/ 5 levels "Camping Equipment",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Product.type     : Factor w/ 21 levels "Binoculars","Climbing Accessories",..: 3 3 3 3 3 3 3 3 3 3
##  $ Product          : Factor w/ 144 levels "Aloe Relief",..: 139 139 139 139 139 139 139 139 139 139
##  $ Order.method.type: Factor w/ 7 levels "E-mail","Fax",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Retailer.country : Factor w/ 21 levels "Australia","Austria",..: 21 5 14 4 12 13 6 16 1 15 ...
##  $ Revenue          : num  315044 13445 NA NA 181120 ...
##  $ Planned.revenue  : num  437477 14313 NA NA 235237 ...
##  $ Product.cost     : num  158372 6299 NA NA 89413 ...
##  $ Quantity         : int  66385 2172 NA NA 35696 NA 15205 7833 NA 14328 ...
##  $ Unit.cost        : num  2.55 2.9 NA NA 2.66 ...
##  $ Unit.price       : num  6.59 6.59 NA NA 6.59 NA 6.59 6.59 NA 6.59 ...
##  $ Gross.profit     : num  156673 7146 NA NA 91707 ...
##  $ Unit.sale.price  : num  5.2 6.19 NA NA 5.49 ...
```

```
summary(rs)
```

```
##       Year                        Product.line
##  Min.   :2004   Camping Equipment       :24108
##  1st Qu.:2005   Golf Equipment          : 8820
```

```
##  Median :2006   Mountaineering Equipment:12348
##  Mean   :2006   Outdoor Protection      : 8820
##  3rd Qu.:2006   Personal Accessories    :30576
##  Max.   :2007
##
##                Product.type                  Product
##  Eyewear             : 9408   Aloe Relief       :  588
##  Watches             : 7644   Astro Pilot       :  588
##  Lanterns            : 7056   Auto Pilot        :  588
##  Cooking Gear        : 5880   Bear Edge         :  588
##  Navigation          : 5880   Bear Survival Edge:  588
##  Climbing Accessories: 4116   Bella             :  588
##  (Other)             :44688   (Other)           :81144
##    Order.method.type  Retailer.country    Revenue
##  E-mail     :12096   Australia: 4032   Min.   :       0
##  Fax        :12096   Austria  : 4032   1st Qu.:   18579
##  Mail       :12096   Belgium  : 4032   Median :   59867
##  Sales visit:12096   Brazil   : 4032   Mean   :  189418
##  Special    :12096   Canada   : 4032   3rd Qu.:  190193
##  Telephone  :12096   China    : 4032   Max.   :10054289
##  Web        :12096   (Other)  :60480   NA's   :59929
##  Planned.revenue     Product.cost        Quantity        Unit.cost
##  Min.   :      16   Min.   :      6   Min.   :     1   Min.   :  0.85
##  1st Qu.:   19557   1st Qu.:   9432   1st Qu.:   328   1st Qu.: 11.43
##  Median :   63907   Median :  32784   Median :  1043   Median : 36.83
##  Mean   :  198818   Mean   : 111625   Mean   :  3607   Mean   : 84.89
##  3rd Qu.:  203996   3rd Qu.: 111371   3rd Qu.:  3288   3rd Qu.: 80.00
##  Max.   :10054289   Max.   :6756853   Max.   :313628   Max.   :690.00
##  NA's   :59929      NA's   :59929     NA's   :59929    NA's   :59929
##    Unit.price       Gross.profit     Unit.sale.price
##  Min.   :   2.06   Min.   : -18160   Min.   :   0.00
##  1st Qu.:  23.00   1st Qu.:   8333   1st Qu.:  20.15
##  Median :  66.77   Median :  25794   Median :  62.65
##  Mean   : 155.99   Mean   :  77793   Mean   : 147.23
##  3rd Qu.: 148.30   3rd Qu.:  78254   3rd Qu.: 140.96
##  Max.   :1359.72   Max.   :3521098   Max.   :1307.80
##  NA's   :59929     NA's   :59929     NA's   :59929
```

```
#split into training and test data
training = rs[(rs$Year == 2004) | (rs$Year == 2005),]
test = rs[(rs$Year == 2006) | (rs$Year == 2007),]
```

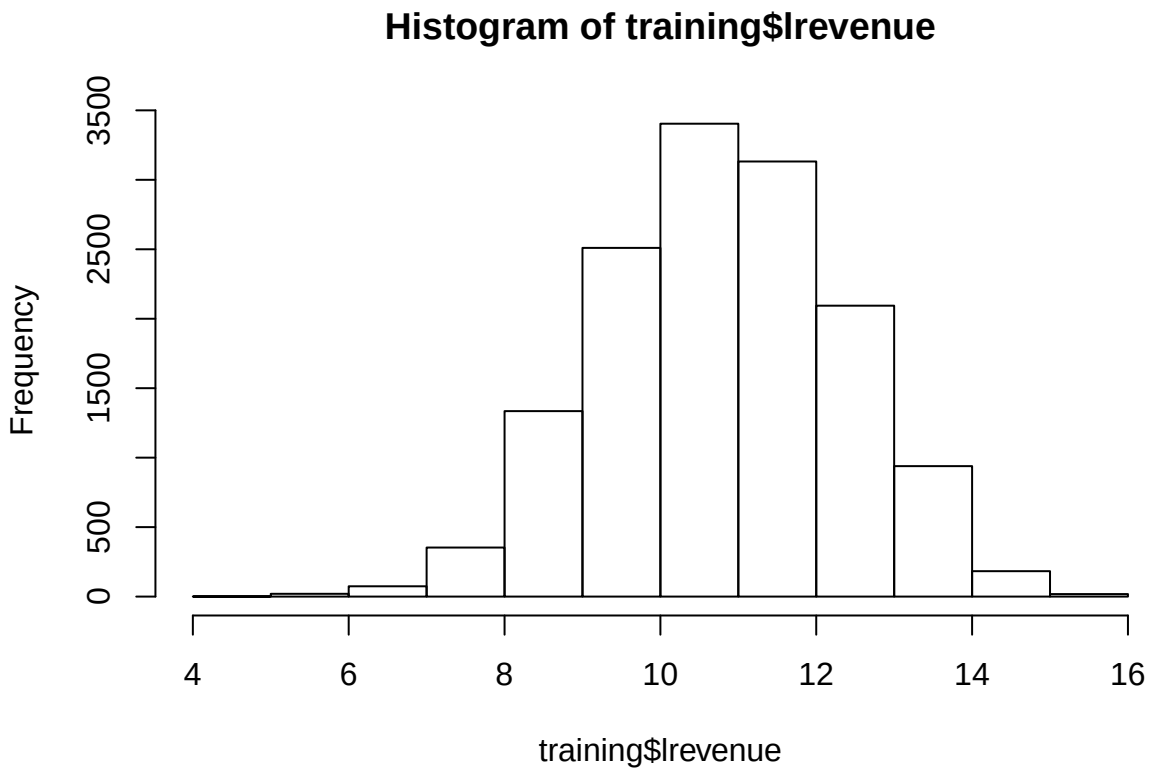
Exploratory data analysis

Use the log of revenue as the dependent variable. this makes it normally distributed. Similar for Product.cost, Unit.price and Unit.cost

```
summary(training$Revenue)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   16690   49640  147200  145200 8390000   28249
```

```
training$Revenue[training$Revenue == 0] = NA
training$lrevenue = log(training$Revenue)
hist(training$lrevenue)
```




```
summary(training$lrevenue)
```

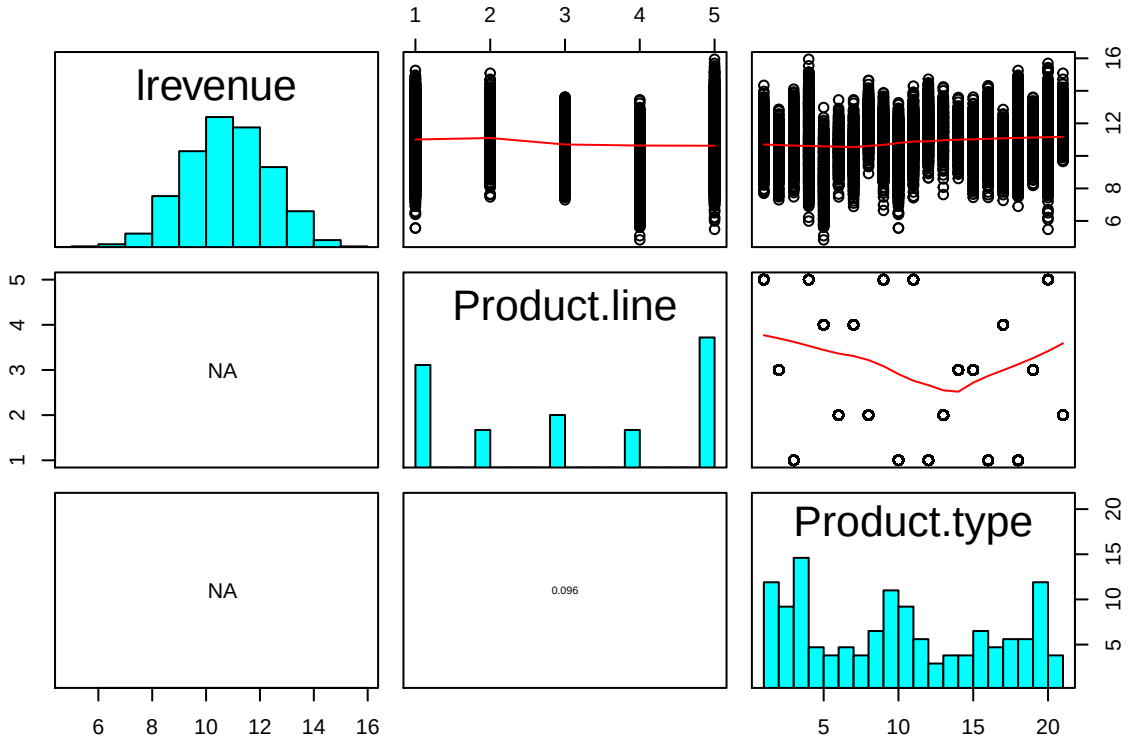
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   4.836   9.730  10.820  10.800  11.890  15.940   28273
```

```
training$lproduct.cost = log(training$Product.cost)
training$lunit.price = log(training$Unit.price)
training$lunit.cost = log(training$Unit.cost)
training$lplanned.revenue = log(training$Planned.revenue)
summary(training$lplanned.revenue)
```

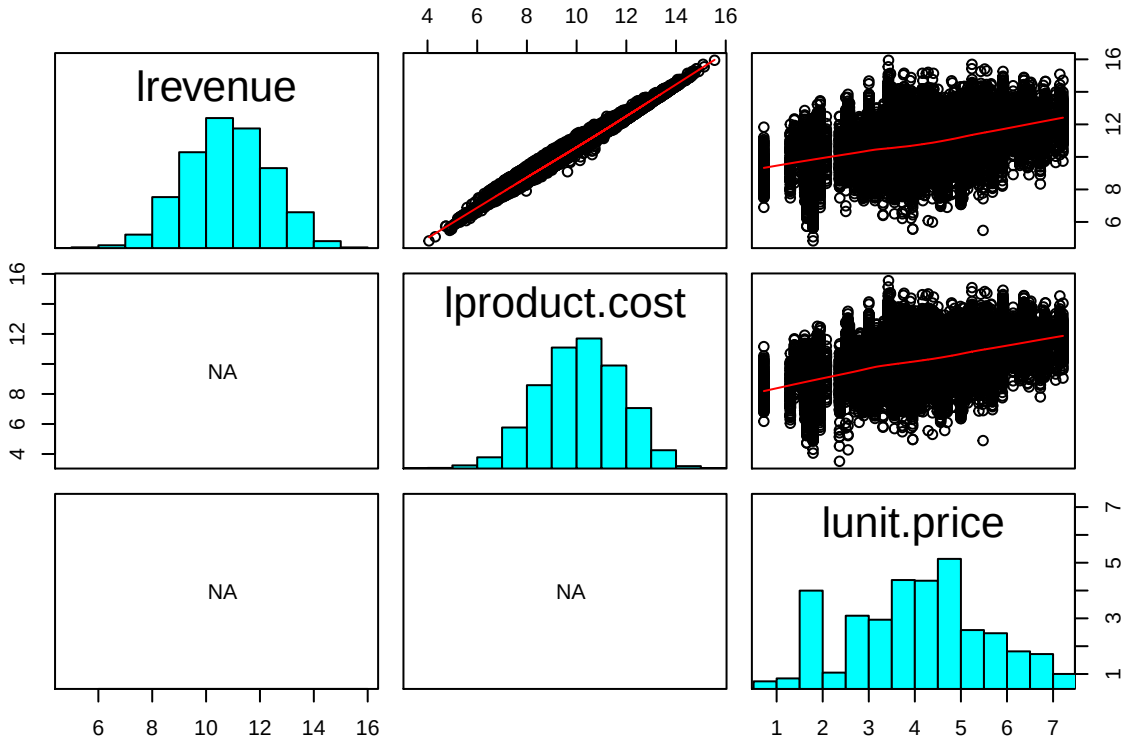
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   4.836   9.774  10.880  10.860  11.960  15.940   28249
```

```
panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
```

```
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(lrevenue~Product.line+ Product.type,data=training, upper.panel=panel.smooth, lower.panel=panel.co
```




```
pairs(lrevenue~lproduct.cost+lunit.price,data=training, upper.panel=panel.smooth, lower.panel=panel.cor
```



It looks like the formula to calculate Revenue is: Revenue = Product.Cost + Gross.profit and and Planned.revenue = Quantity * Unit.Price

```
model1 = lm(lrevenue~lplanned.revenue+lproduct.cost+Product.line+Product.line*lproduct.cost, data=train
summary(model1)
```

```
##
## Call:
## lm(formula = lrevenue ~ lplanned.revenue + lproduct.cost + Product.line +
##     Product.line * lproduct.cost, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25523 -0.01288  0.01069  0.03029  0.10749
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                    0.0858167  0.0067929
## lplanned.revenue                               0.9265258  0.0029168
## lproduct.cost                                  0.0626493  0.0028212
## Product.lineGolf Equipment                    -0.0136699  0.0132367
## Product.lineMountaineering Equipment          -0.0150954  0.0156772
## Product.lineOutdoor Protection                 0.0421336  0.0108167
## Product.linePersonal Accessories              -0.0849514  0.0081731
## lproduct.cost:Product.lineGolf Equipment       0.0013826  0.0012151
```

```
## lproduct.cost:Product.lineMountaineering Equipment  0.0029689  0.0015189
## lproduct.cost:Product.lineOutdoor Protection        0.0002759  0.0011850
## lproduct.cost:Product.linePersonal Accessories      0.0111835  0.0007792
##                                                    t value Pr(>|t|)
## (Intercept)                                         12.633  < 2e-16 ***
## lplanned.revenue                                   317.648  < 2e-16 ***
## lproduct.cost                                       22.206  < 2e-16 ***
## Product.lineGolf Equipment                          -1.033   0.3018
## Product.lineMountaineering Equipment                -0.963   0.3356
## Product.lineOutdoor Protection                       3.895 9.86e-05 ***
## Product.linePersonal Accessories                   -10.394  < 2e-16 ***
## lproduct.cost:Product.lineGolf Equipment             1.138   0.2552
## lproduct.cost:Product.lineMountaineering Equipment   1.955   0.0506 .
## lproduct.cost:Product.lineOutdoor Protection         0.233   0.8159
## lproduct.cost:Product.linePersonal Accessories      14.352  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06037 on 14052 degrees of freedom
##   (28273 observations deleted due to missingness)
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9985
## F-statistic: 9.171e+05 on 10 and 14052 DF,  p-value: < 2.2e-16
```

```
model2 = lm(Revenue~Planned.revenue+Product.cost+Product.line+Product.line*Product.cost, data=training)
summary(model2)
```

```
##
## Call:
## lm(formula = Revenue ~ Planned.revenue + Product.cost + Product.line +
##     Product.line * Product.cost, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -355044    -385    1275    2258   84041
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -1.121e+03  1.508e+02
## Planned.revenue                                   8.966e-01  2.244e-03
## Product.cost                                      6.359e-02  3.574e-03
## Product.lineGolf Equipment                       -2.925e+02  3.241e+02
## Product.lineMountaineering Equipment              1.628e+03  3.680e+02
## Product.lineOutdoor Protection                    1.667e+03  2.977e+02
## Product.linePersonal Accessories                 -1.067e+03  2.107e+02
## Product.cost:Product.lineGolf Equipment          -1.648e-02  1.910e-03
## Product.cost:Product.lineMountaineering Equipment  1.123e-02  3.935e-03
## Product.cost:Product.lineOutdoor Protection        4.001e-02  1.005e-02
## Product.cost:Product.linePersonal Accessories      1.025e-01  9.177e-04
##                                                  t value Pr(>|t|)
## (Intercept)                                       -7.439 1.07e-13 ***
## Planned.revenue                                  399.530  < 2e-16 ***
## Product.cost                                      17.792  < 2e-16 ***
## Product.lineGolf Equipment                        -0.902  0.36684
## Product.lineMountaineering Equipment               4.424 9.77e-06 ***
```

```
## Product.lineOutdoor Protection                       5.601 2.17e-08 ***
## Product.linePersonal Accessories                    -5.063 4.18e-07 ***
## Product.cost:Product.lineGolf Equipment             -8.627  < 2e-16 ***
## Product.cost:Product.lineMountaineering Equipment    2.854  0.00432 **
## Product.cost:Product.lineOutdoor Protection          3.983 6.85e-05 ***
## Product.cost:Product.linePersonal Accessories      111.649  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9238 on 14052 degrees of freedom
##   (28273 observations deleted due to missingness)
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.9991
## F-statistic: 1.537e+06 on 10 and 14052 DF,  p-value: < 2.2e-16
```

** Is the change in the average revenue different from 95 cents when the planned revenue increases by $1?**

```
linearHypothesis(model2, "Planned.revenue = 0.95",vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Planned.revenue = 0.95
##
## Model 1: restricted model
## Model 2: Revenue ~ Planned.revenue + Product.cost + Product.line + Product.line *
##     Product.cost
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1  14053
## 2  14052  1 19.246 1.158e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result is highly statistically significant, supporting the hypothesis that the change in average revenuw is 95 cents when planned revenue increases by $1