Lab 2

Ron Cordell, Lei Yang, Subhashini Raghunathan

```
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
```

```
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 ...
```

```r
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

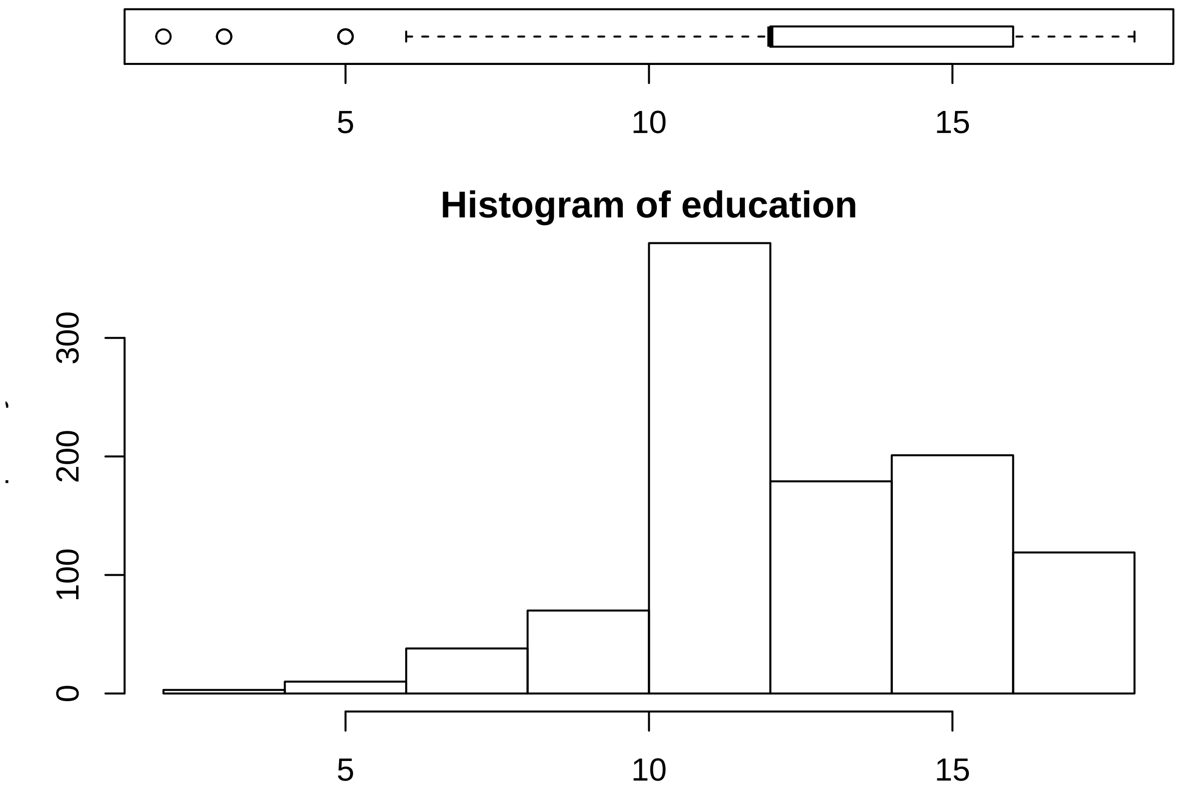
```r
summary(education)
```

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

```r
str(education)
```

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

```r
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

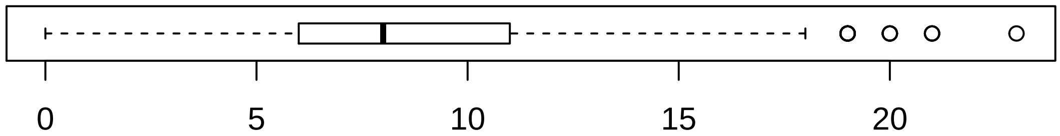
```
summary(experience)
```

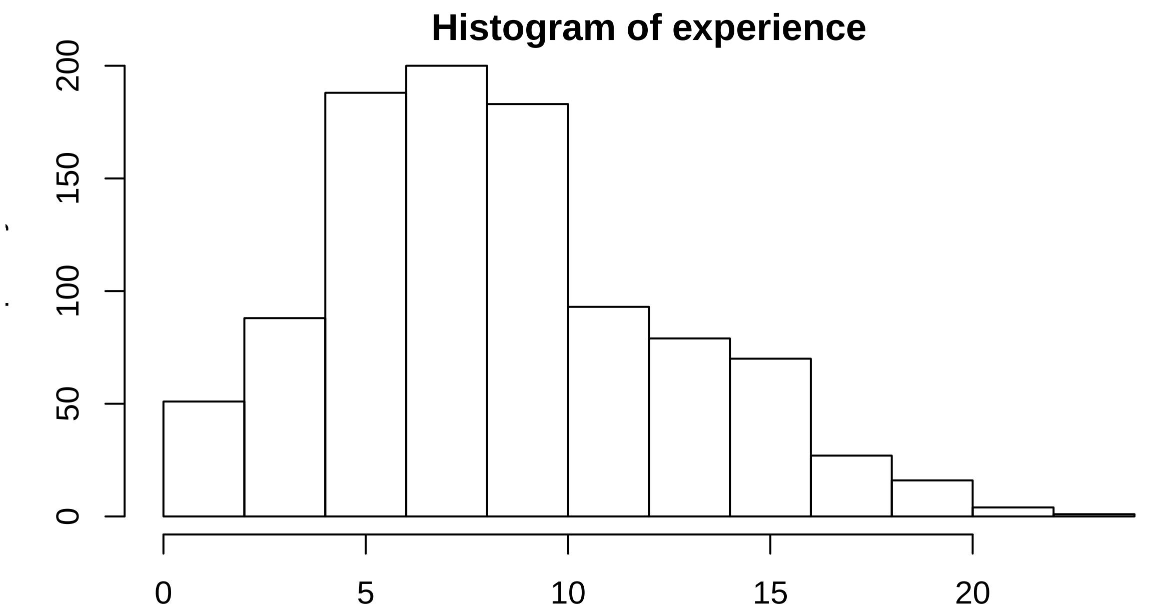
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.000   8.000   8.788  11.000  23.000
```

```
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)
```





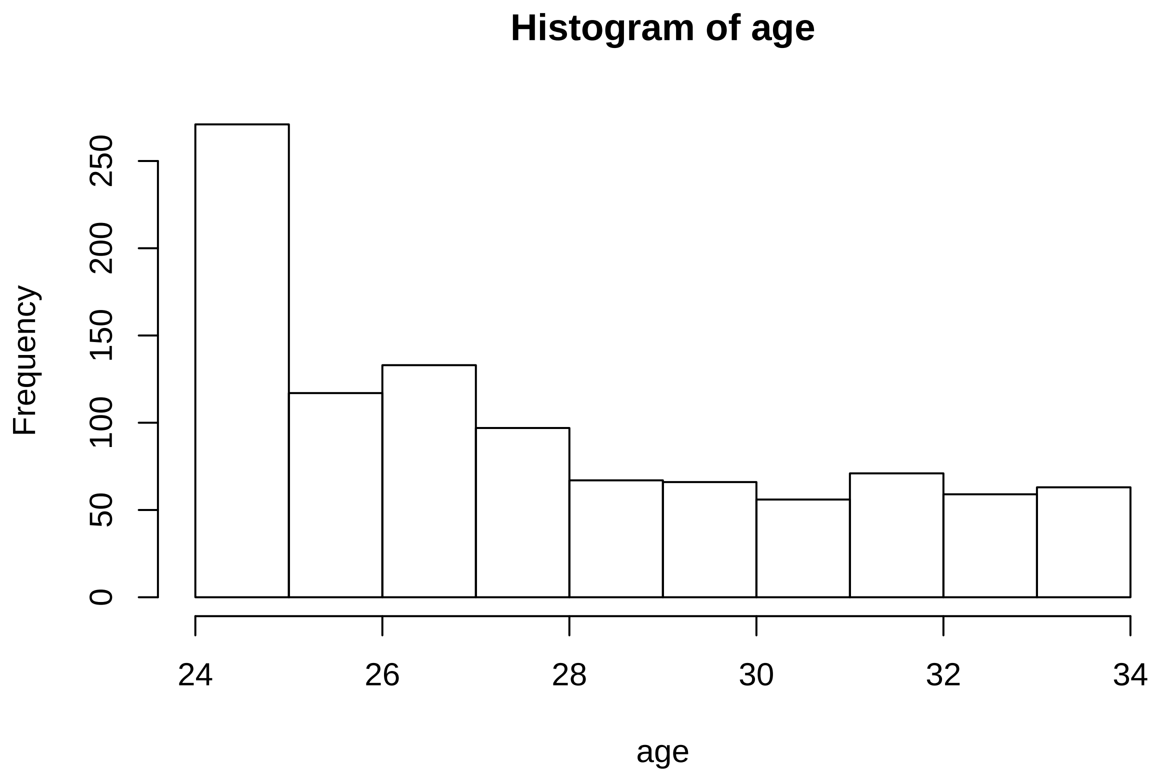

```r
summary(age)
```

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

```r
str(age)
```

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

```r
hist(age)
```



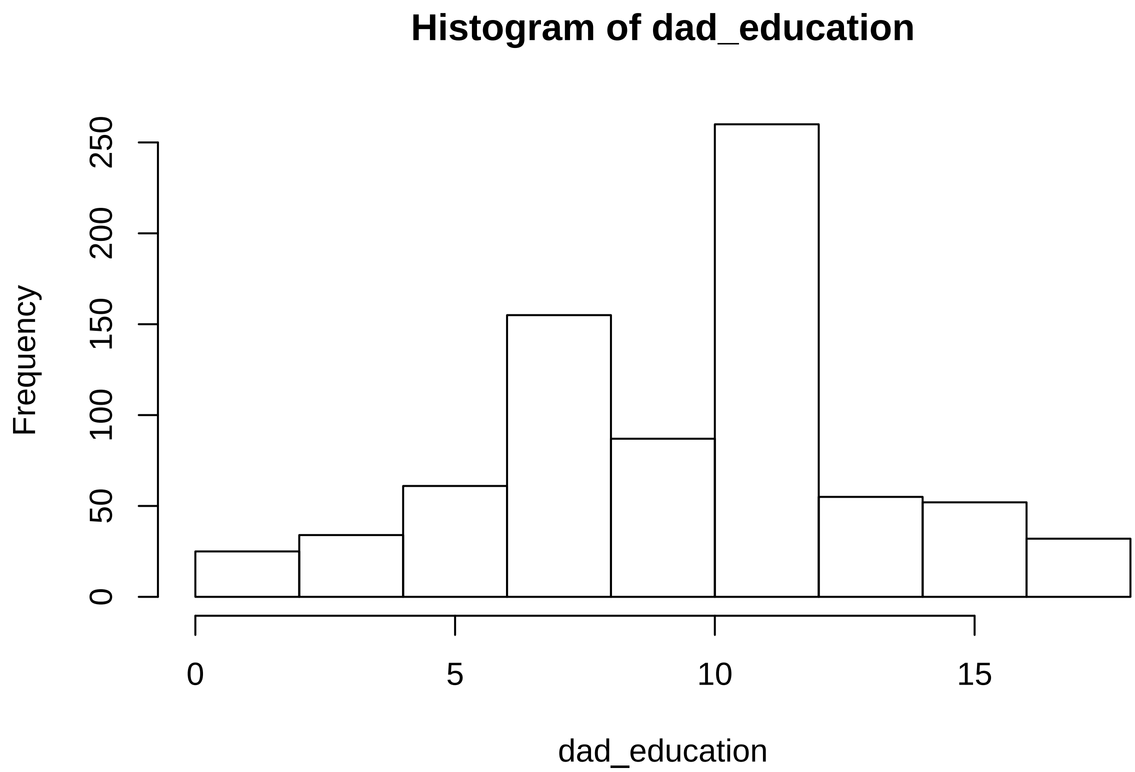

```r
summary(dad_education)
```

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

```r
str(dad_education)
```

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

```r
hist(dad_education)
```




```r
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
```

```r
str(mom_education)
```

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

```r
hist(mom_education)
```

Has quite a few missing observations (316)

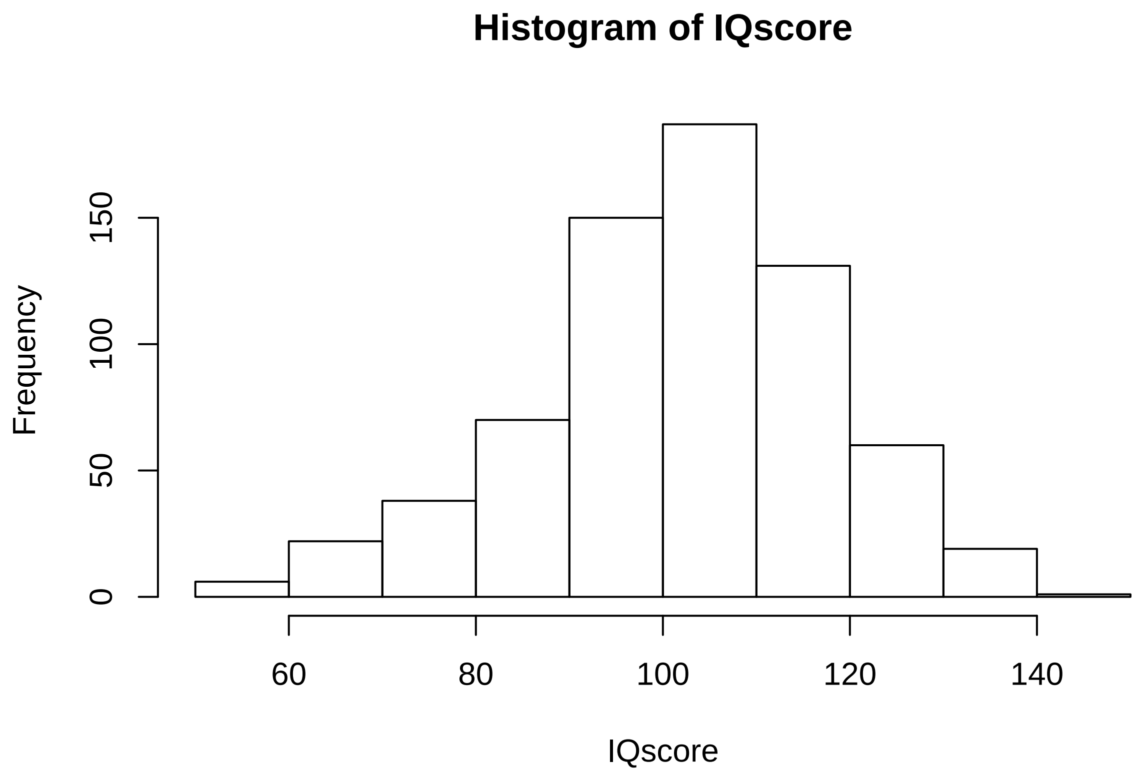
```r
summary(IQscore)
```

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

```r
str(IQscore)
```

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

```r
hist(IQscore)
```



almost normal distribution

```r
summary(logWage)
```

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

```r
str(logWage)
```

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

```r
hist(logWage)
```


```r
summary(raceColor)
```

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

```r
table(raceColor)
```

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

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

```r
summary(city)
```

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

```r
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
```

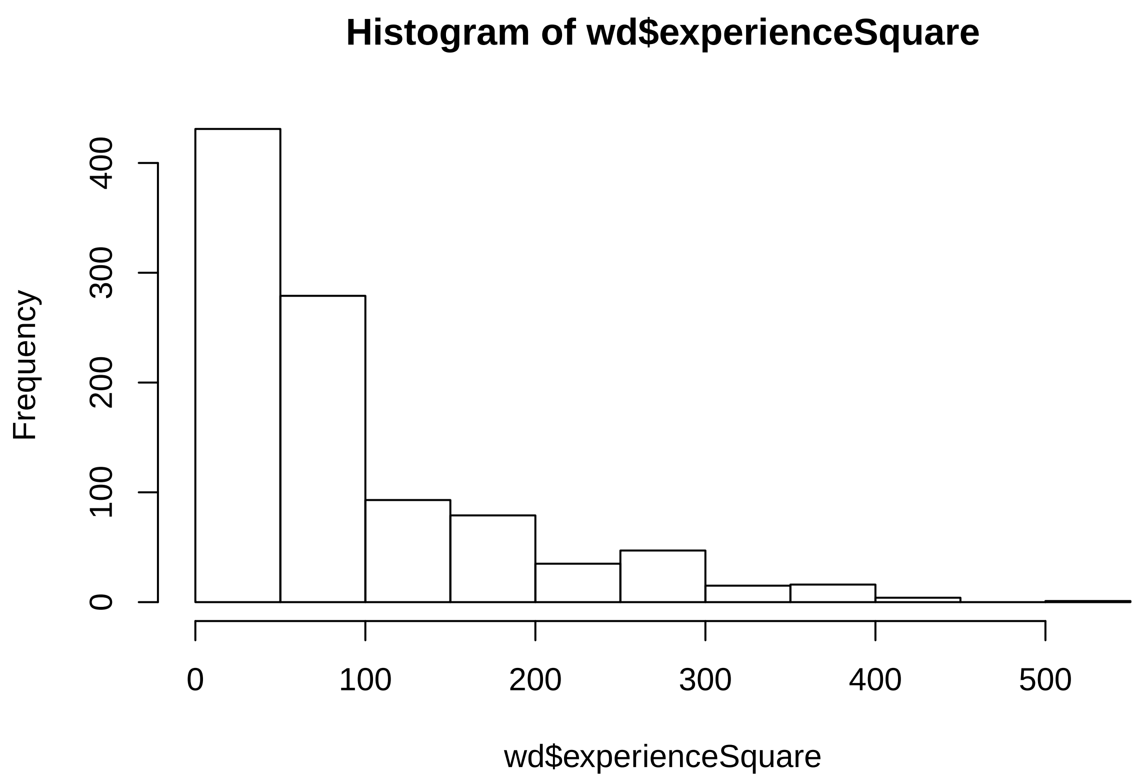
```r
table(z1)
```

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

```r
table(z2)
```

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

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

hist(wd$experienceSquare)
```



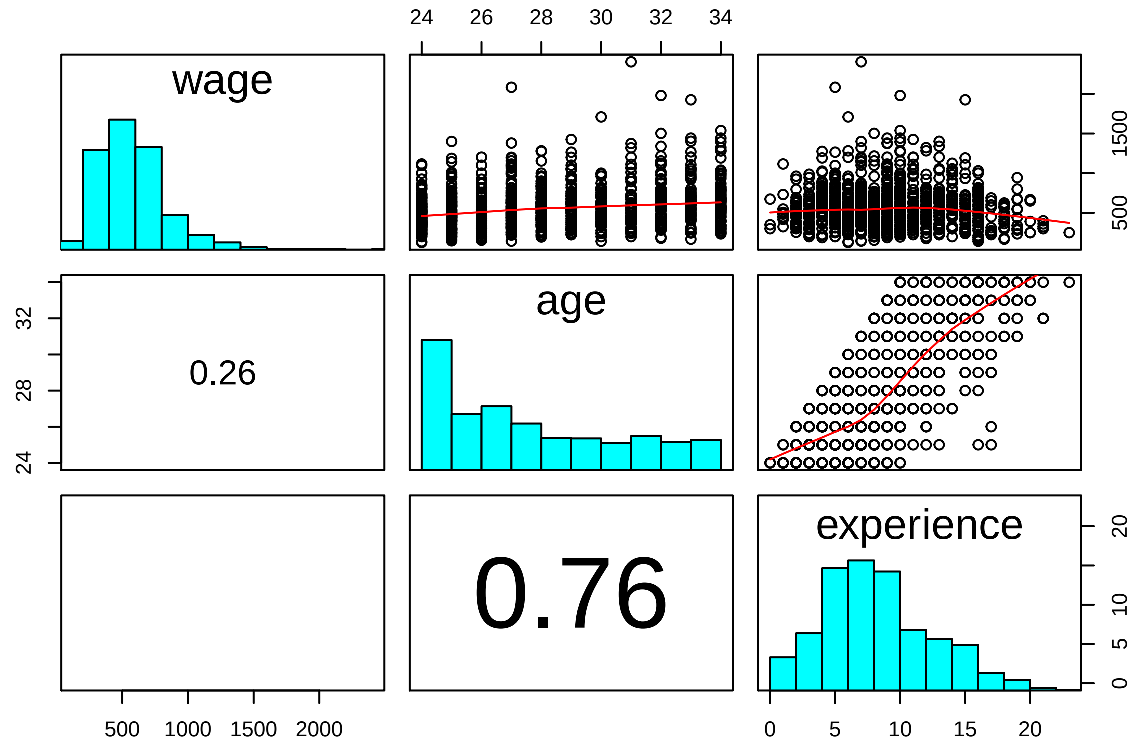

```
attach(wd)
```

```
## 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
```

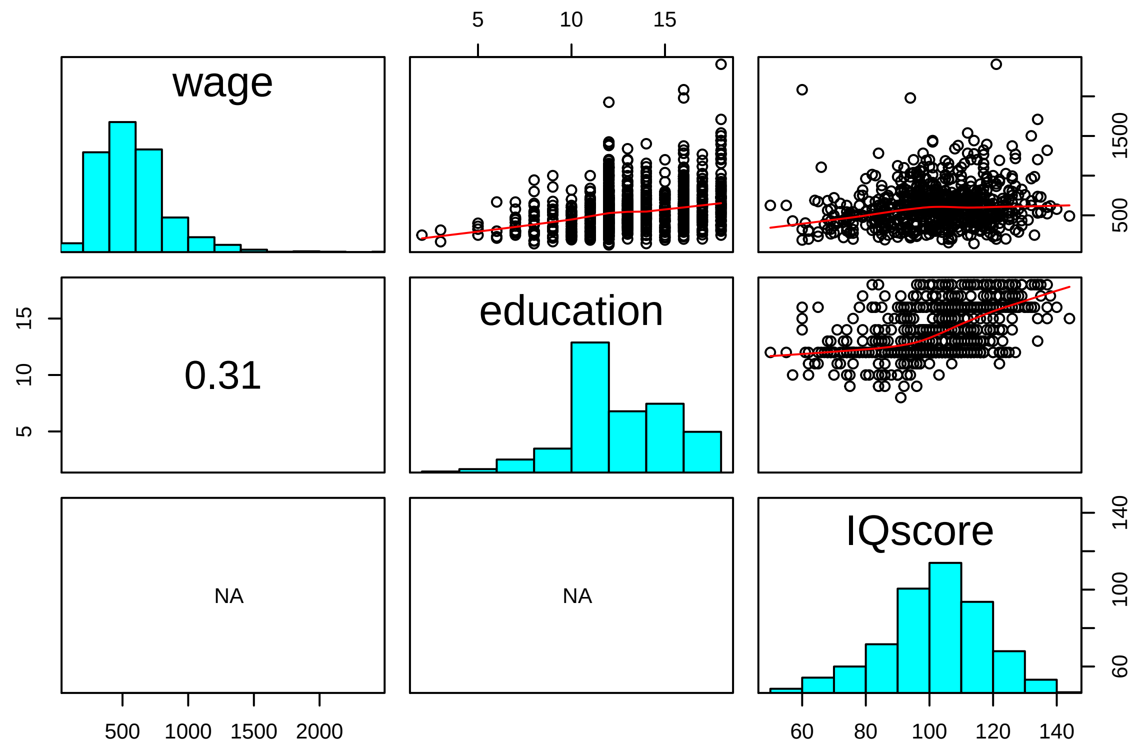
4.2 bivariate analysis

```
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)
```

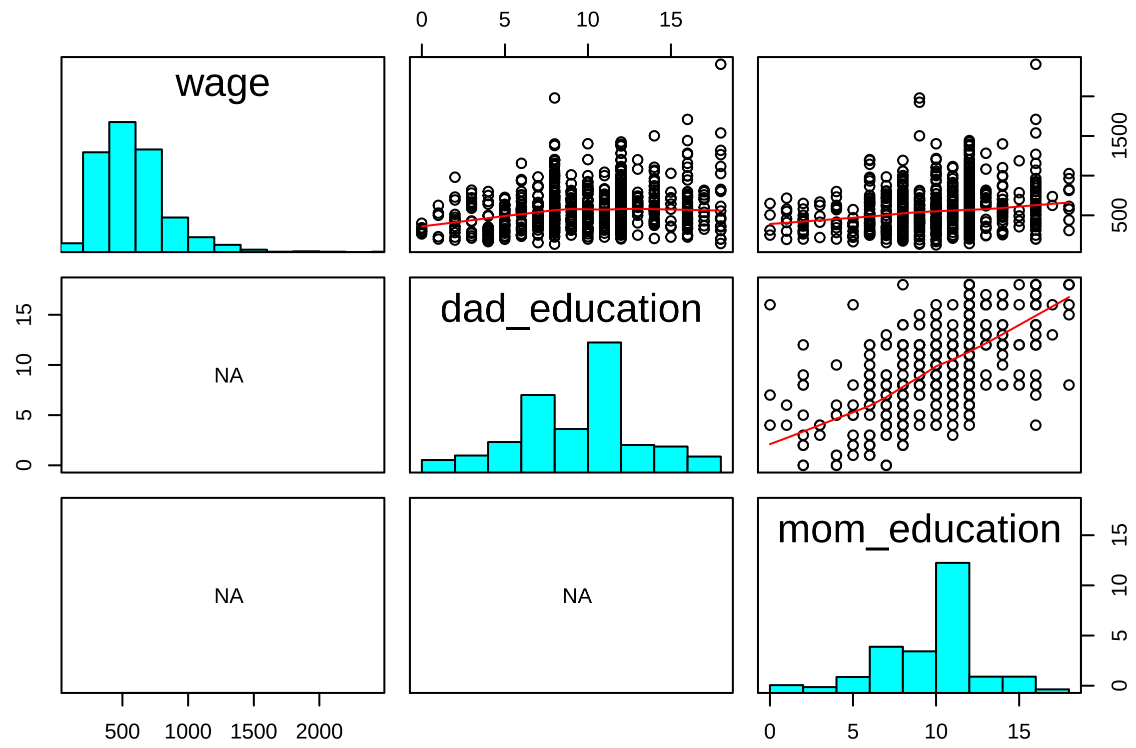
```
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

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



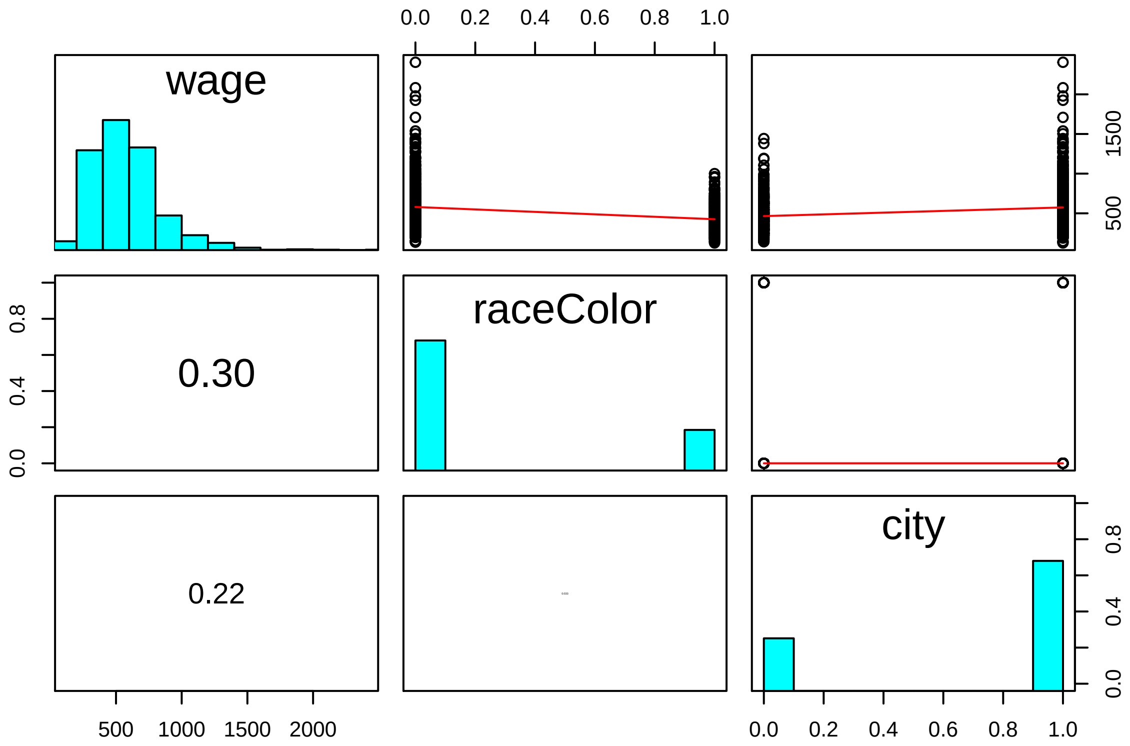

```
pairs(wage~education+IQscore,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.panel=panel
```



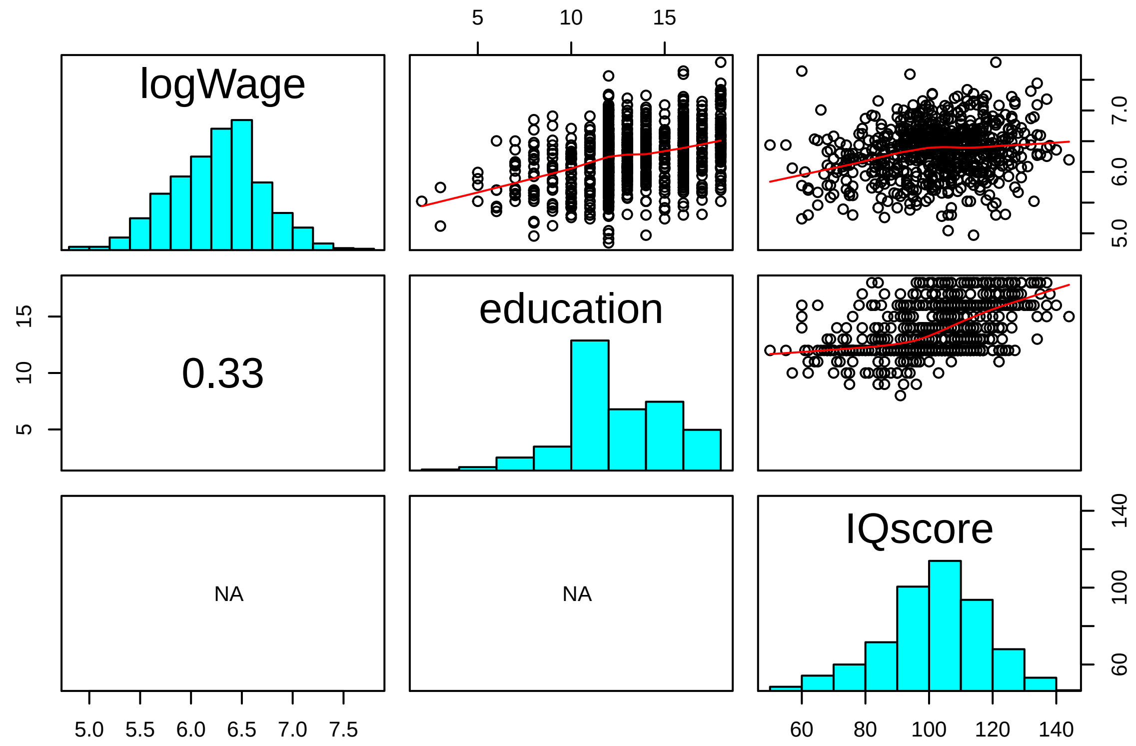

```
pairs(wage~dad_education+mom_education,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.pa
```



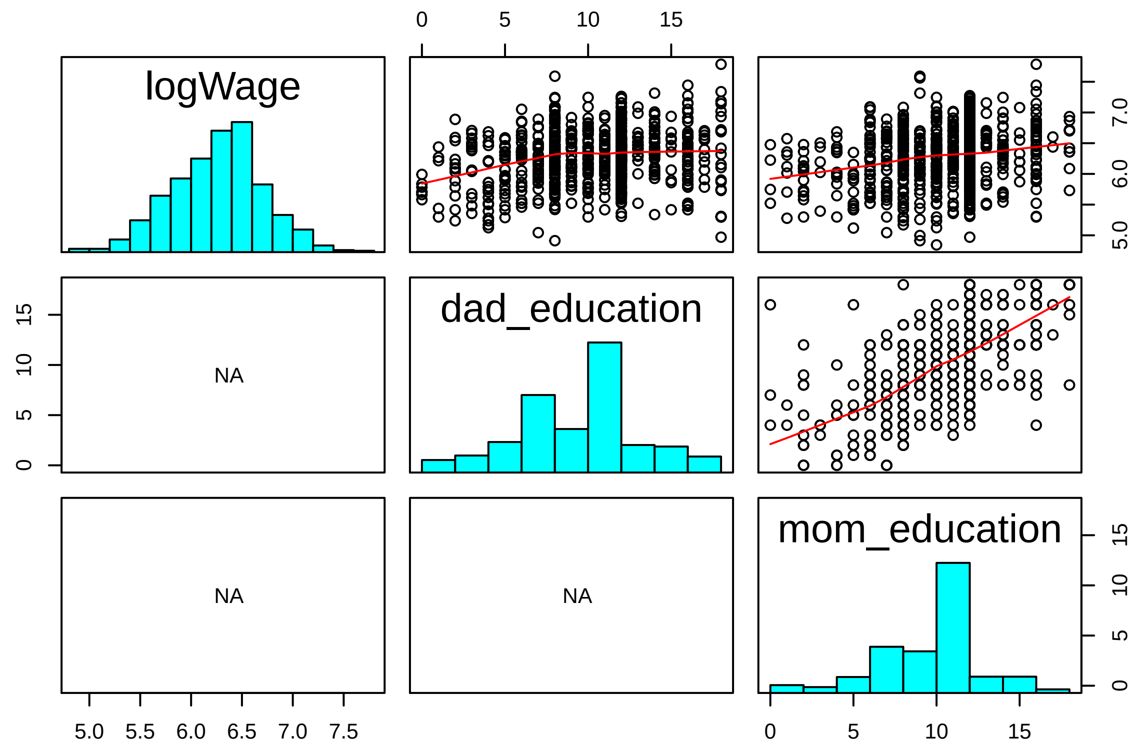

```
pairs(wage~raceColor+city,data=wd, upper.panel=panel.smooth, lower.panel=panel.cor, diag.panel=panel.hi
```




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




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

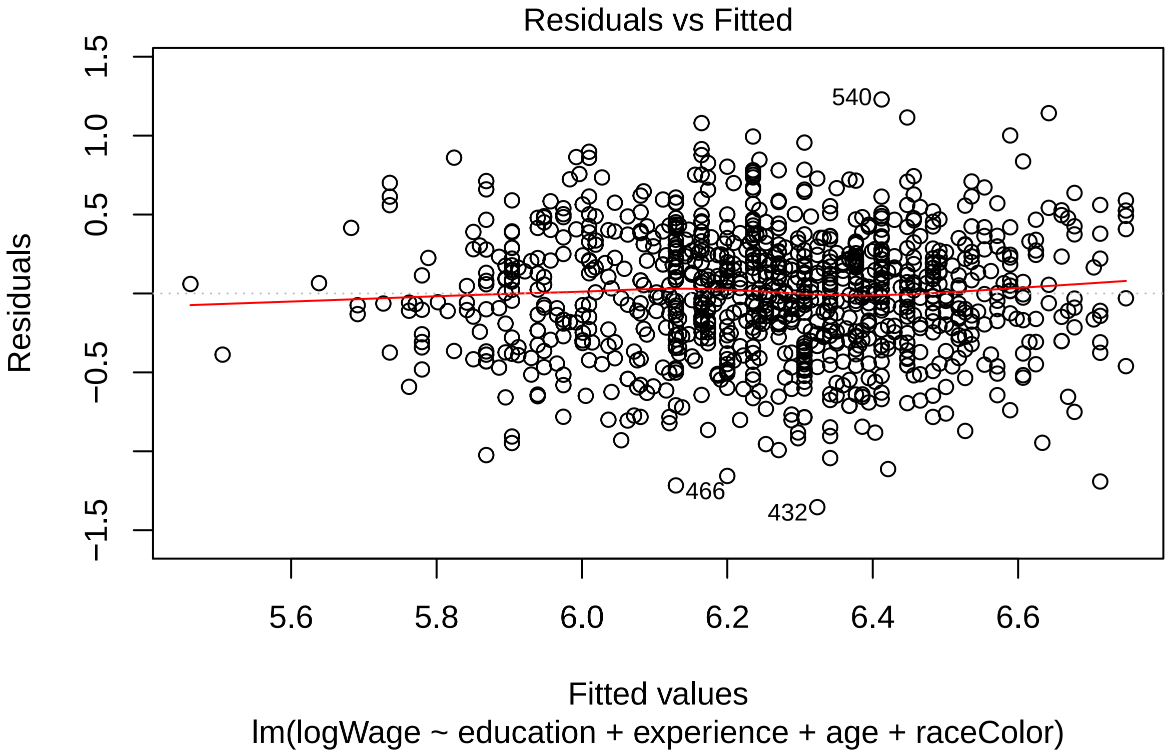


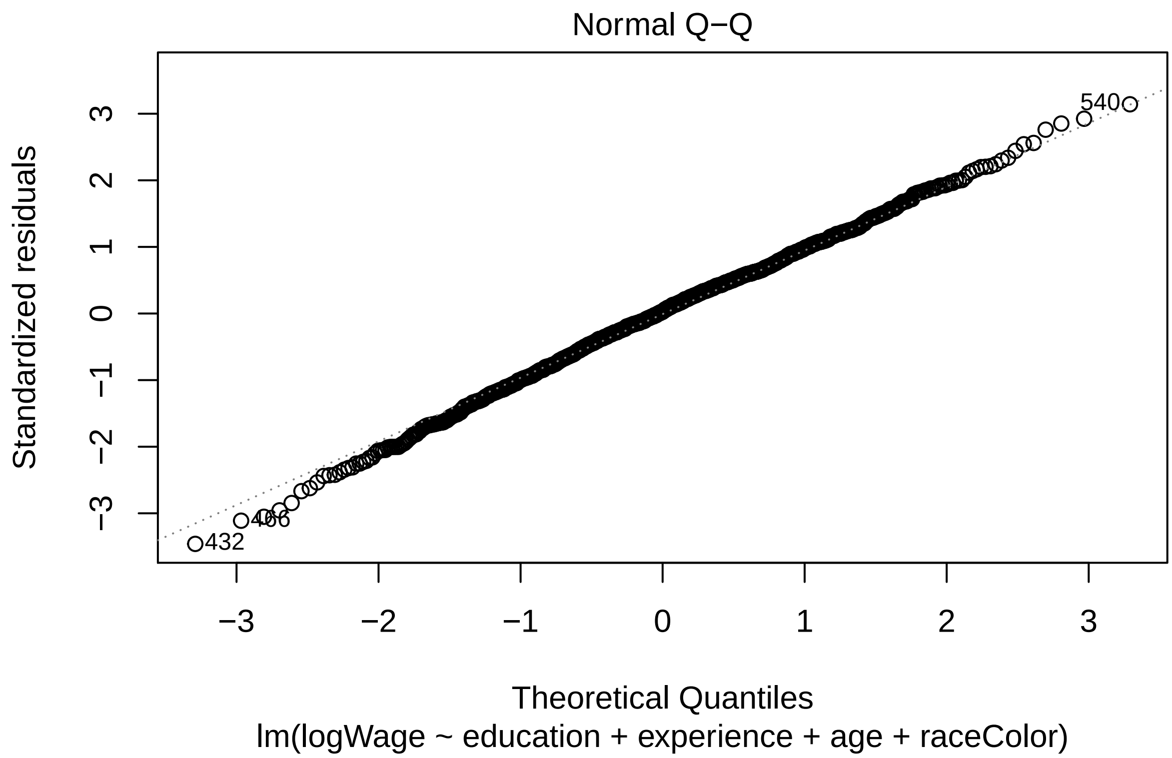
4.3

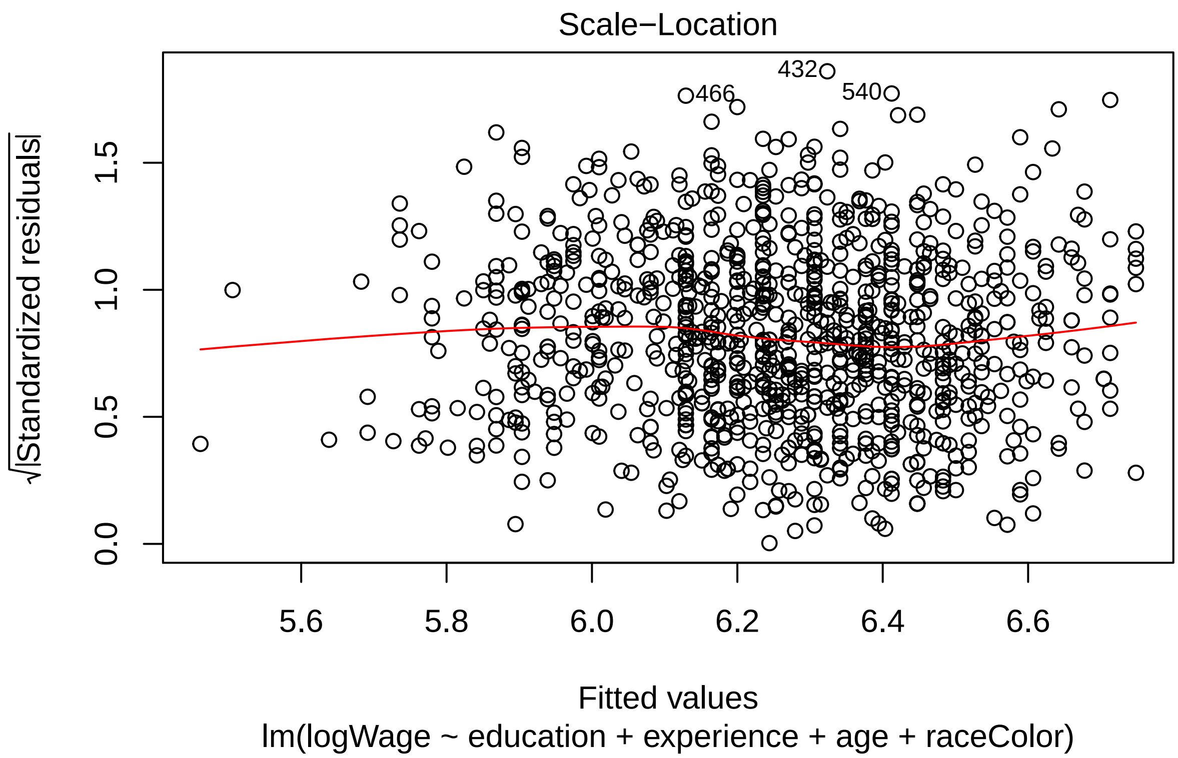
```r
model1 = lm(logWage~education+experience+age+raceColor, data=wd)
coeftest(model1)
```

```
##
## 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
```

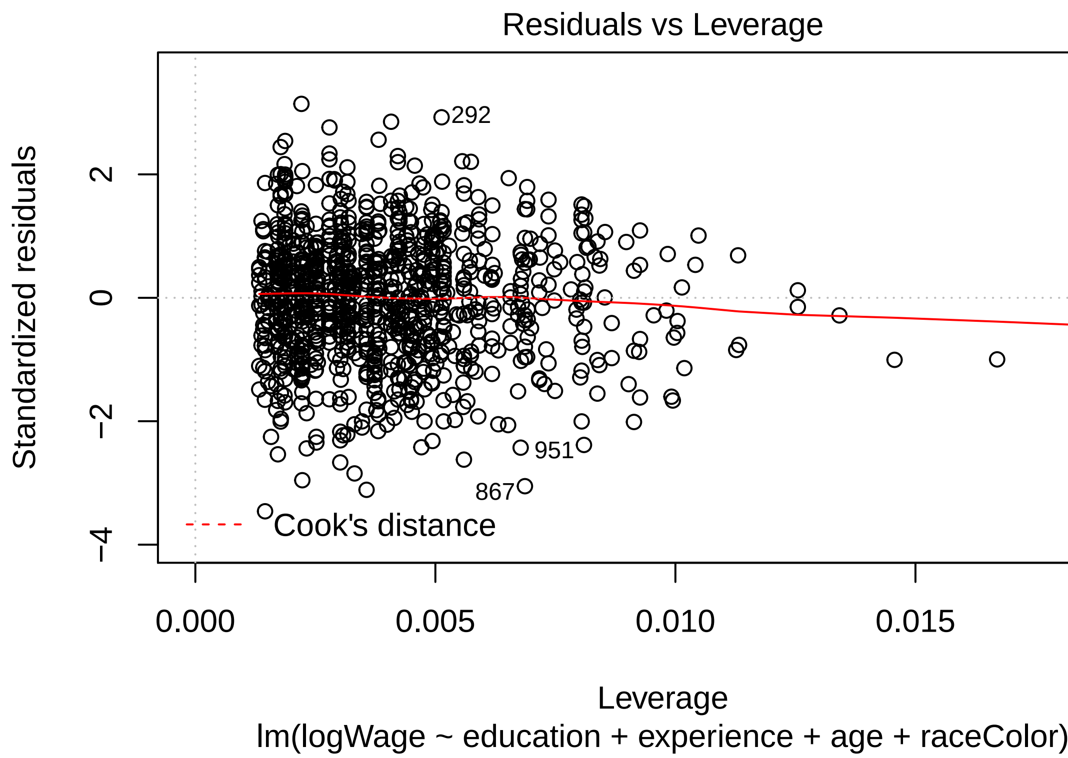
```r
plot(model1)
```


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


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(logWage ~ education + experience + age + raceColor)
540
432
406
−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)




```r
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.

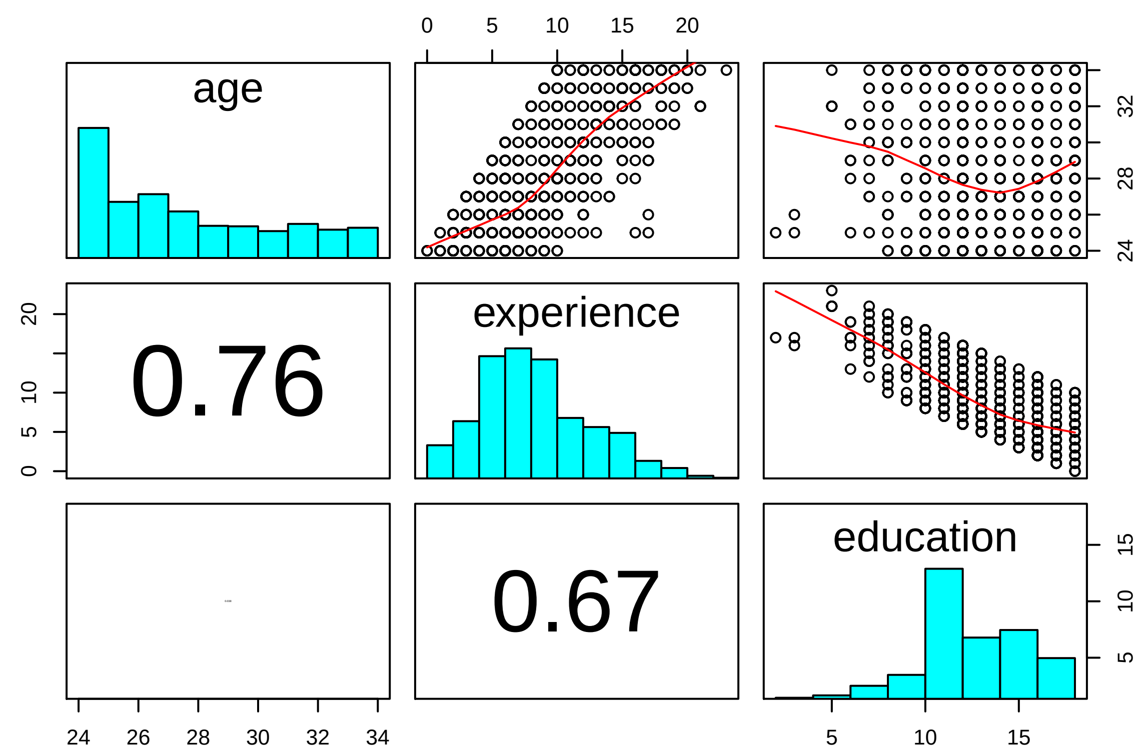
Using the summary function to display parameters (not necessary to use the heteroskedasticity-robust versions here)

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 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=panel
```




```
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
```

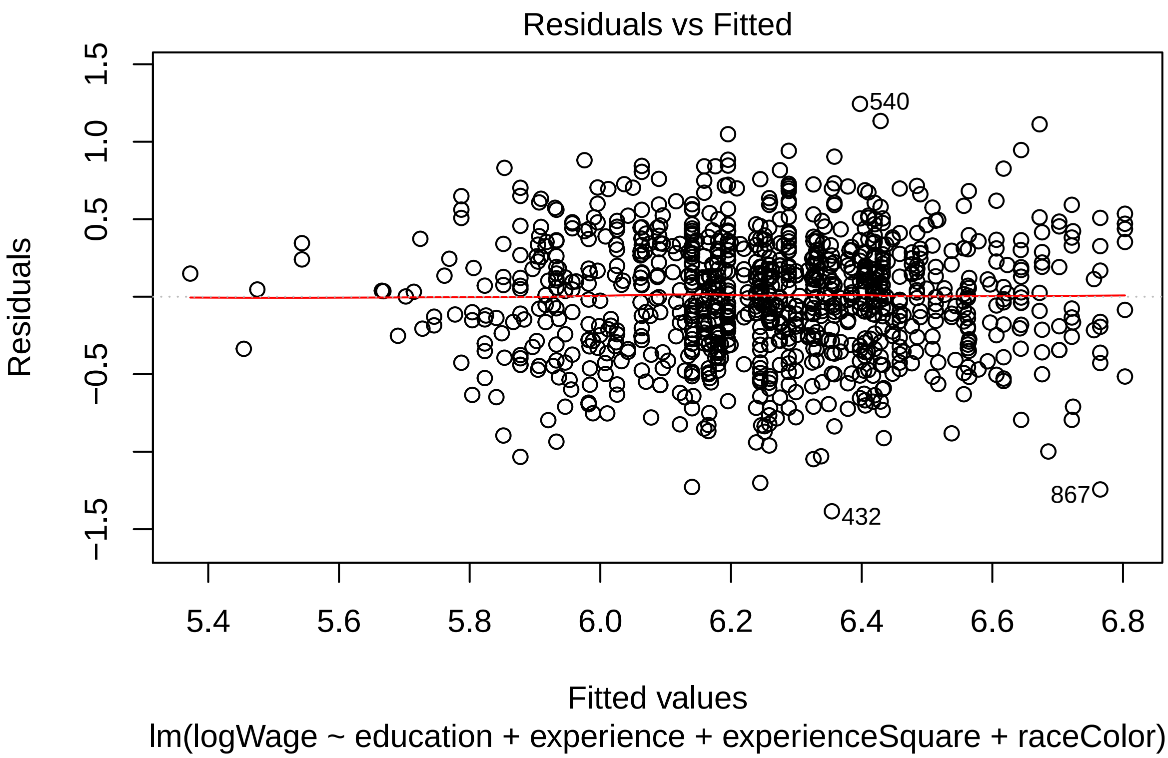
```
## -1.35396 -0.25550  0.01074  0.24867  1.22932
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.961661   0.113346  43.774   <2e-16 ***
## education    0.079608   0.006376  12.486   <2e-16 ***
## experience   0.035372   0.003988   8.869   <2e-16 ***
## raceColor   -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
```

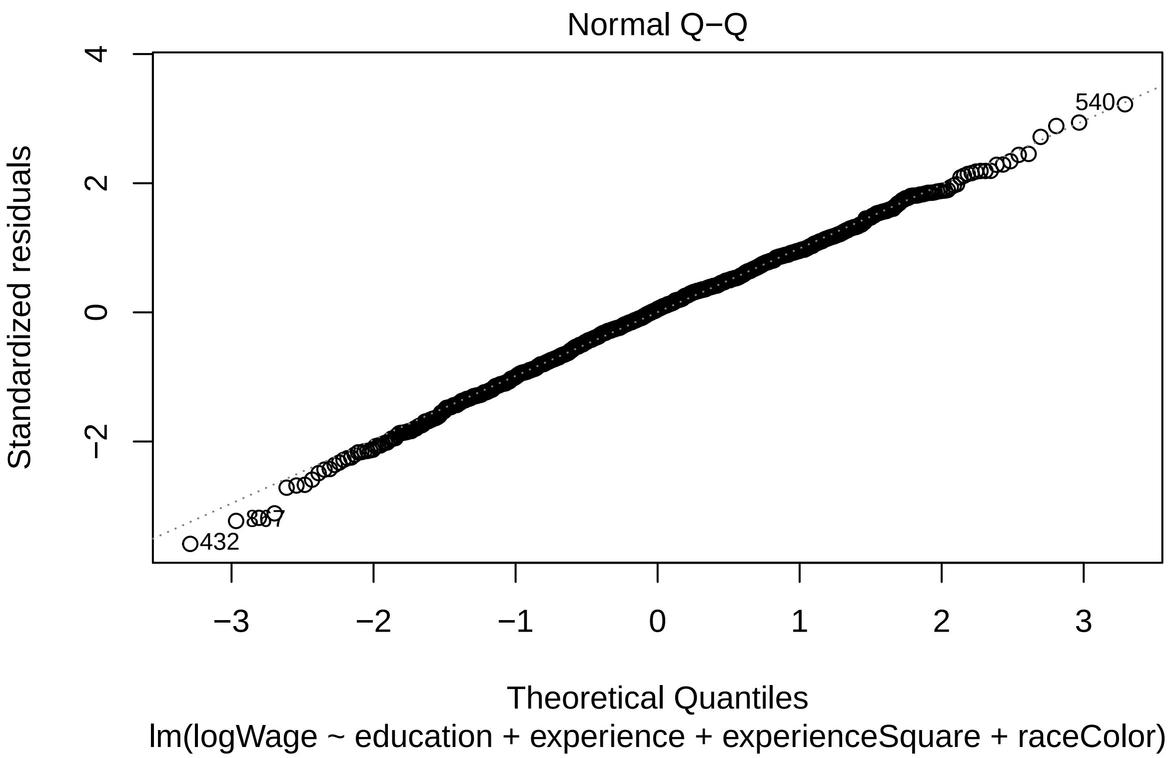
The coeff on education is ~ 0.08, meaning that an increase in 1 year of education leads to an 8% increase in wages, holding experience and raceColor fixed.

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.

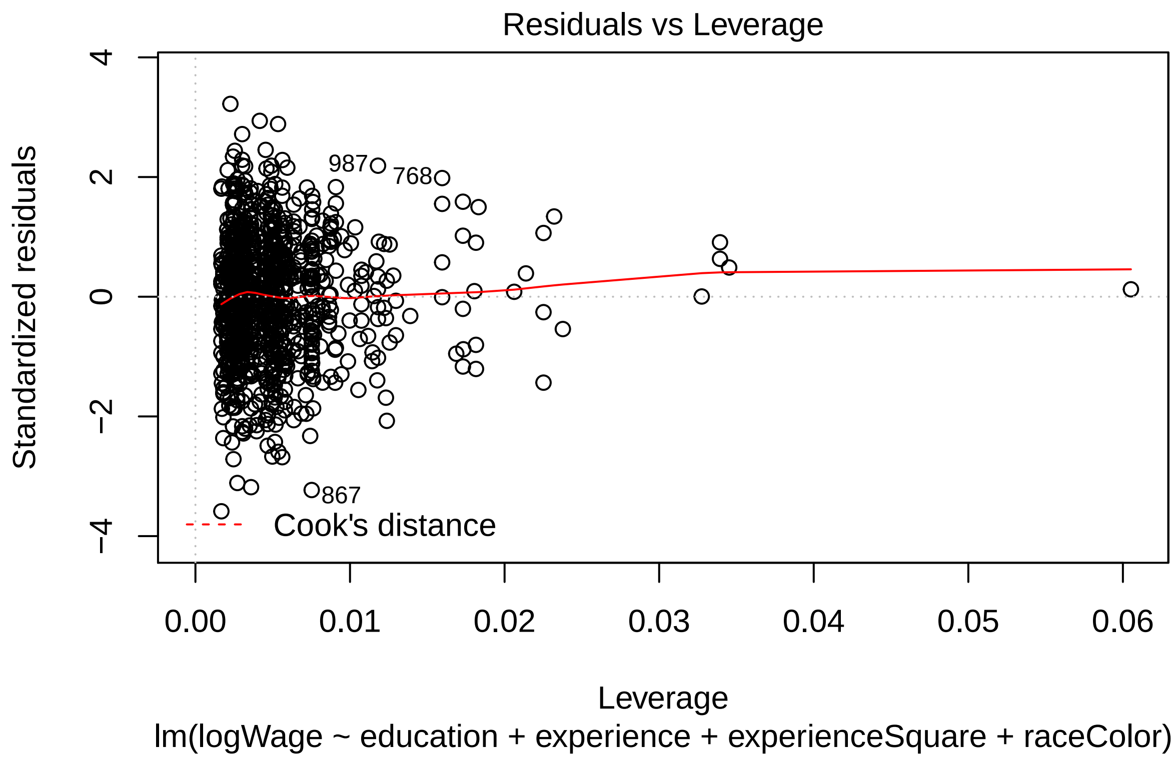
4.4

```
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
432
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)


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

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




```r
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
```

```r
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
##
```

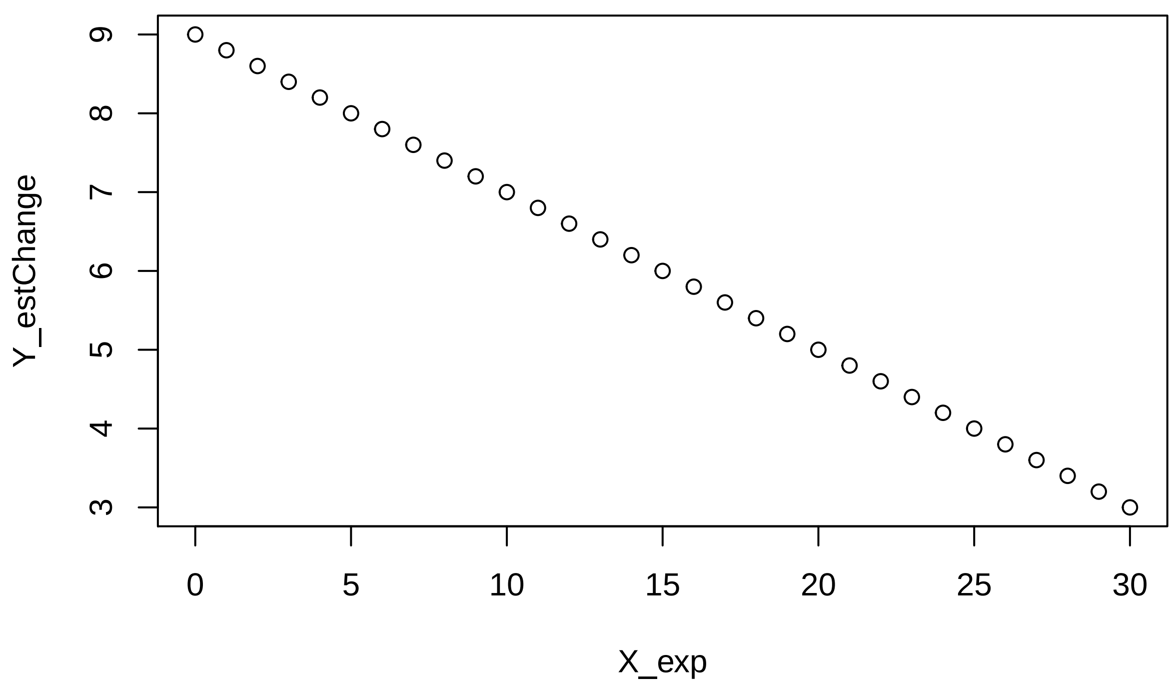
```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.7355175  0.1197719  39.538  < 2e-16 ***
## education         0.0794641  0.0062917  12.630  < 2e-16 ***
## experience        0.0924930  0.0115147   8.033 2.68e-15 ***
## experienceSquare -0.0028779  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

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

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



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

4.5

```
model4 = lm(logWage~education+experience+experienceSquare+raceColor+dad_education+mom_education+rural+ci
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
```

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
```

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

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

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

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

```
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

4.5.2:

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

4.5.3

```
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)
```

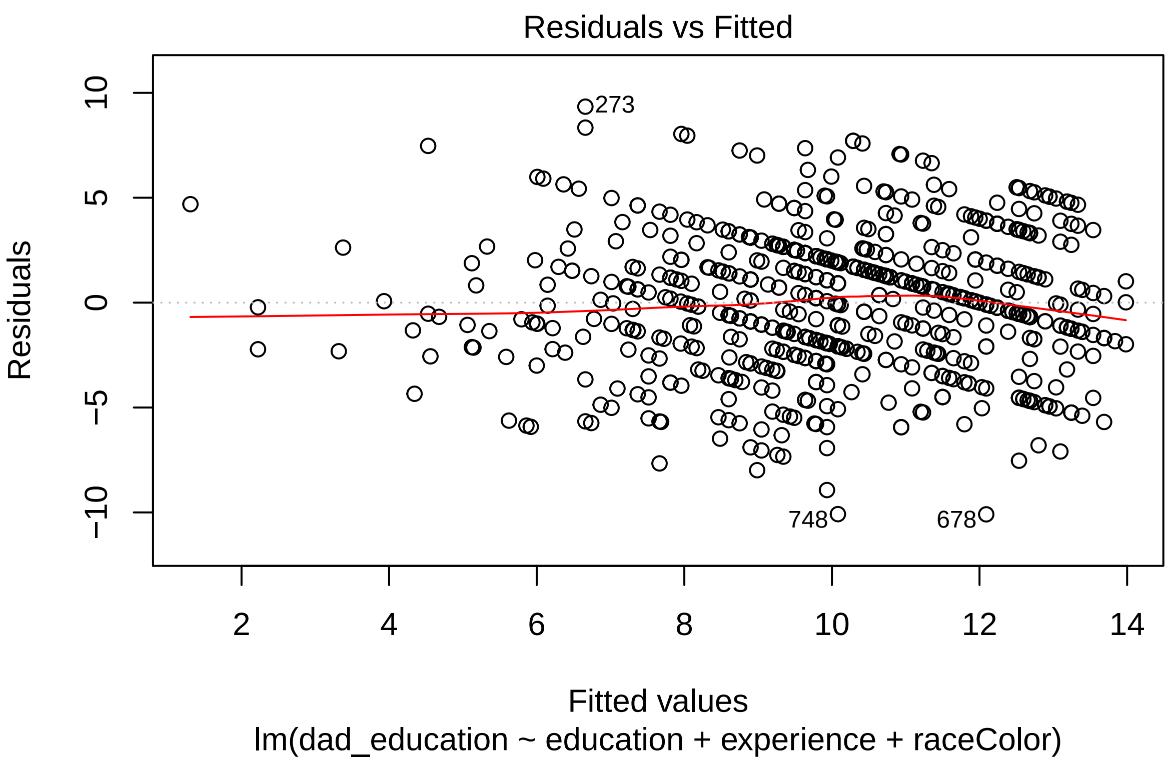
```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_educ2 + mom_educ2 + rural + city, data = wd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30741 -0.23286  0.01943  0.24786  1.28807
##
## Coefficients:
##                    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
##
```

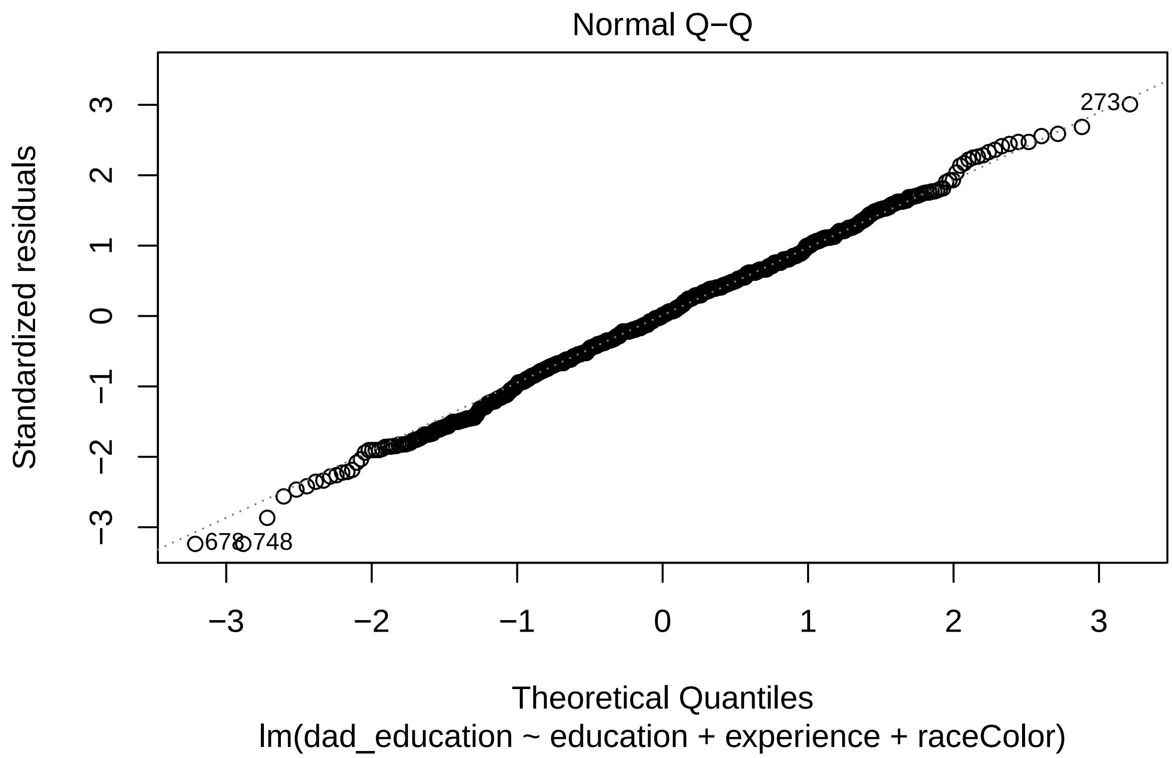
```
## Residual standard error: 0.3764 on 991 degrees of freedom
## Multiple R-squared:  0.2981, Adjusted R-squared:  0.2925
## F-statistic: 52.62 on 8 and 991 DF,  p-value: < 2.2e-16
```

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

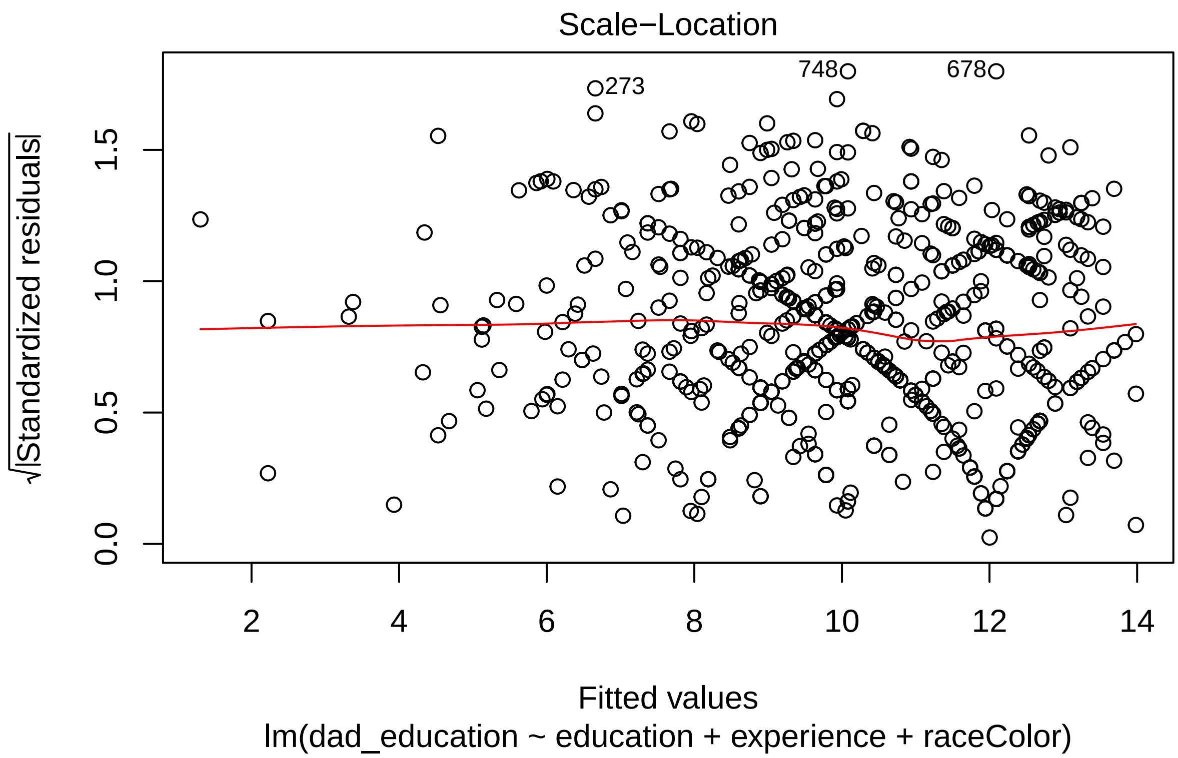
4.5.4

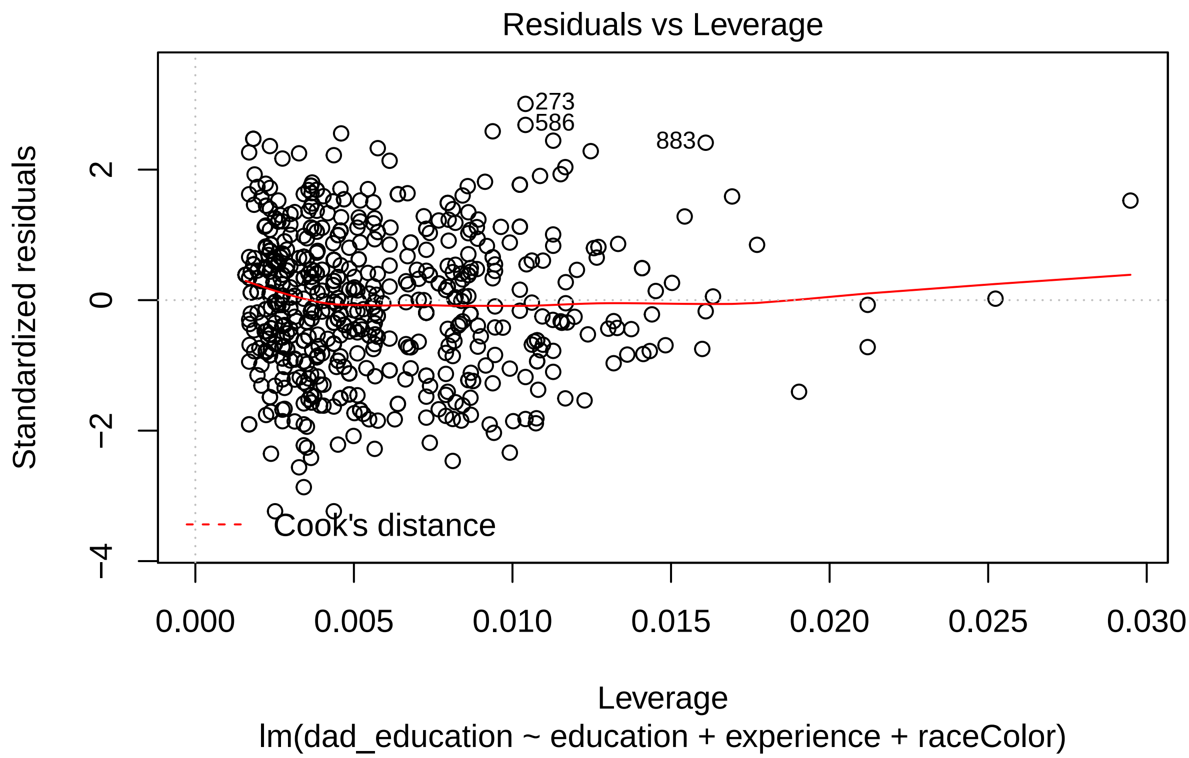
```
model6 =lm(dad_education~education+experience+raceColor, data=wd)
plot(model6)
```





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


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)




```r
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$experience$ - 2.12raceColor

```
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$rac
sum(is.na(wd$dad_educ3))
```

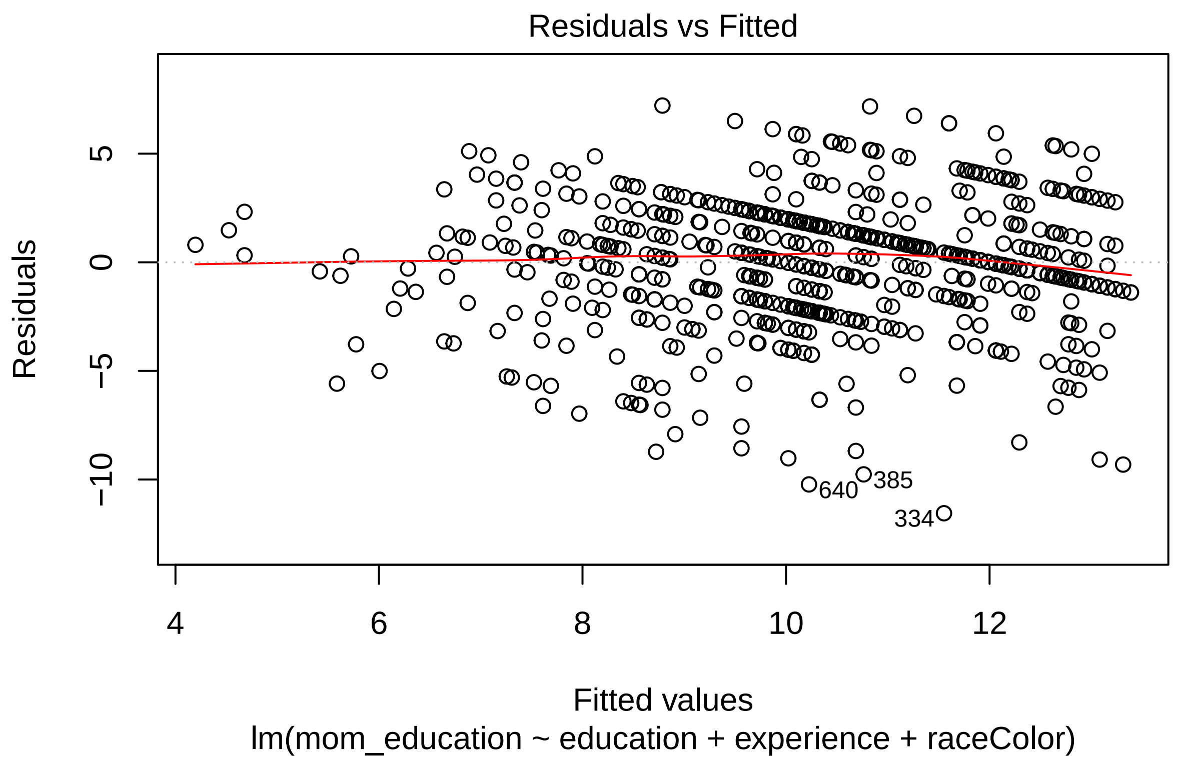
```
## [1] 239
```

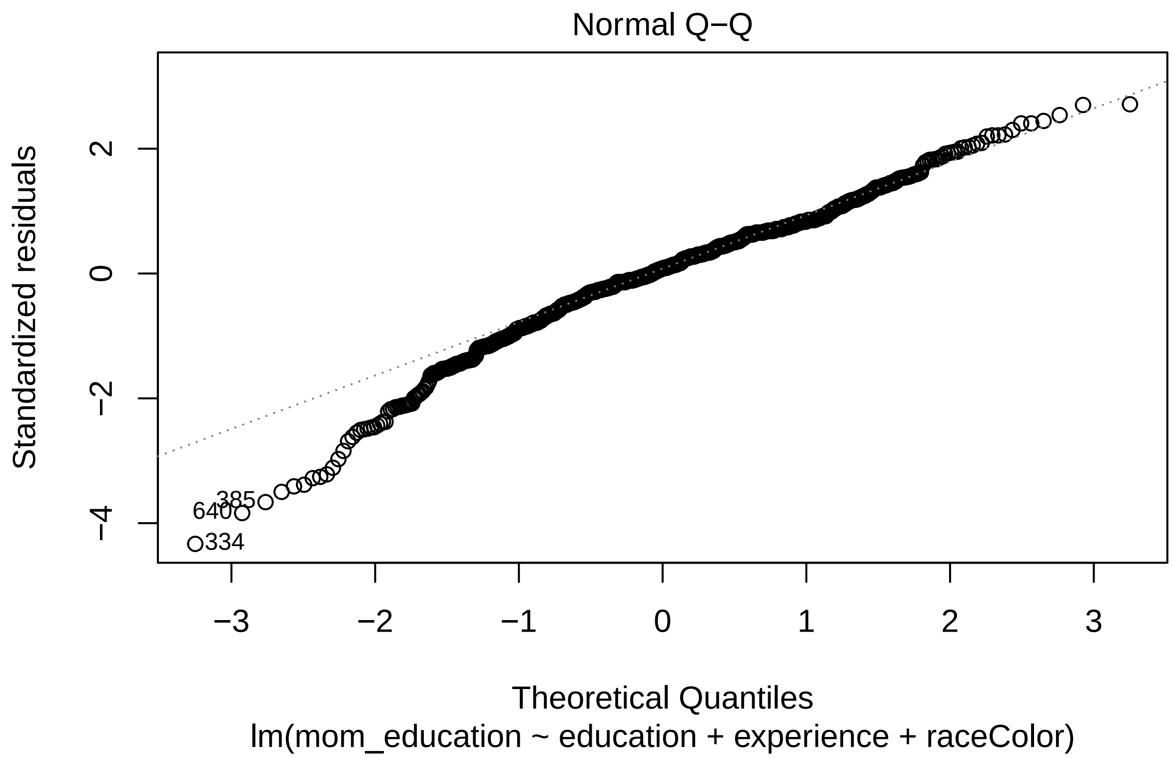
```
sum(is.na(wd_to_fix$dad_educ3))
```

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

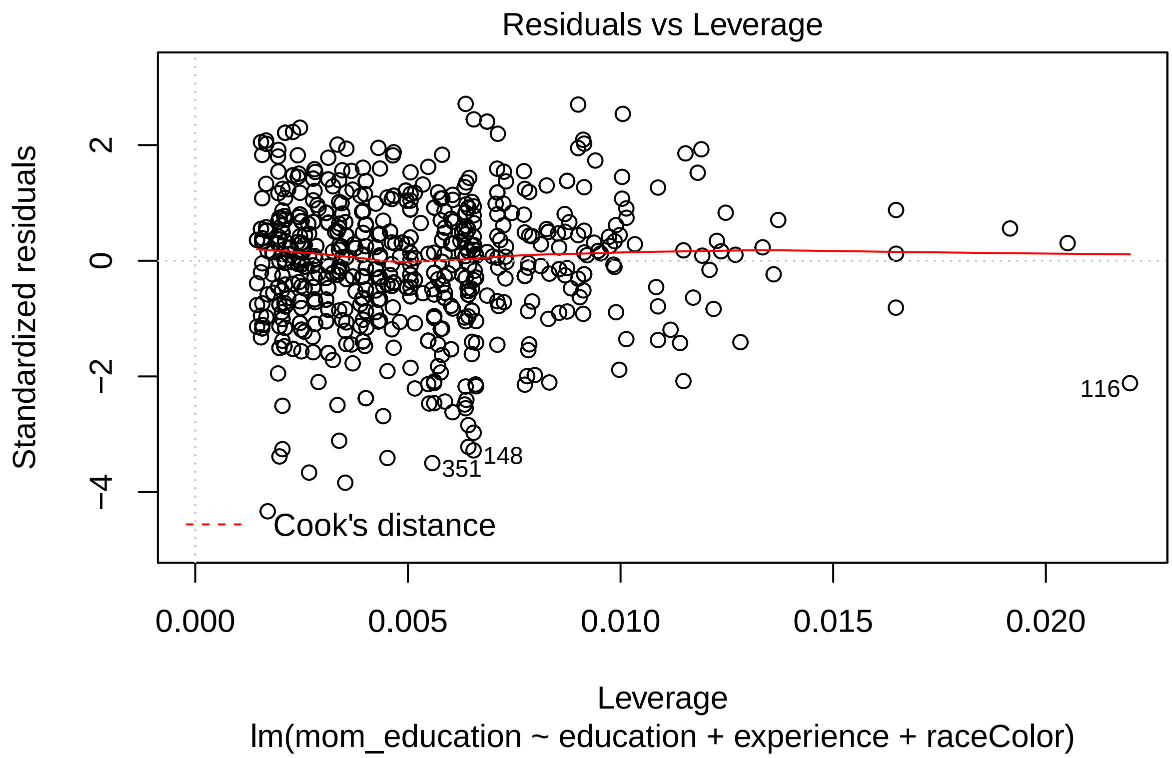
```
wd$dad_educ3[is.na(wd$dad_educ3)] = wd_to_fix$dad_educ3
```

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




```r
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

```r
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
```

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

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

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

```r
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.

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

4.6.1

Z1 must be uncorrelated with the error term u

4.6.2