W271 Lab2

Dr. Who

February 27, 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, $E(Y|X)$.

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & x < 0 \text{ or } x > 1 \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$E(X) = \int_0^1 xdx = \frac{1}{2}x^2\Big|_0^1 = \frac{1}{2} - 0 = \frac{1}{2}$$

Now we take the left part of the ruler, assuming the ruler starts at 0 and the left half has length $E(X)$. Breaking the left part of the ruler at position Y which has a uniform probability distribution:

$$f(y) = \begin{cases} \frac{1}{E(X)}, & 0 \le y \le E(X) \\ 0, & 1y < 0 \text{ or } y > E(X) \end{cases}$$

$$E(Y) = \int_{-\infty}^{\infty} yf(y)dy$$

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

Therefore, since $E(X) = \frac{1}{2}$ and since $E(Y)$ is conditional on X to begin with:

$$E(Y\big|X) = \frac{1}{2}E(X)$$

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.

$$E(Y) = E(E(Y\big|X))$$

$$E(Y\big|X) = \frac{1}{2}E(X) \text{ therefore } E(Y) = E(\frac{1}{2}E(X))$$

$$\text{Since } E(X) = \frac{1}{2} \text{ we have: } E(Y) = E(\frac{1}{2} \times \frac{1}{2}) = \frac{1}{4}$$

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

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{x}, & 0 \leq y \leq x \\ 0, & y < 0 \text{ or } y > 1 \end{cases}$$

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

The conditional probability function of X given Y is given by the joint probability density function divided by the marginal probability density function:

$$f_X(x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

$$f_Y(y) = \begin{cases} \frac{1}{a}, & 0 \leq y \leq a \\ 0, & y < 0 \text{ or } y > a \end{cases}$$

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

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.

Question 3: Turtles

Next, suppose that the lifespan of a species of turtle follows a uniform distribution 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 $y_1, y_2, ..., y_n$.

$$f(y|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq y \leq \theta \\ 0, & y < 0 \text{ or } y > \theta \end{cases}$$

$$L(\theta) = \prod_{i=1}^{n} f(y;\theta) = \begin{cases} \theta^{-n}, & 0 \leq y_i \leq \theta \\ 0, & y_i < 0 \text{ or } y_i > \theta \end{cases}$$

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

Since $L(\theta) = \theta^{-n}$ for $0 \leq y_i \leq \theta$ then it follows that $\theta \geq y_i$ for all i.

Therefore the MLE for θ is $\hat{\theta} = max(x_1, x_2, ..., x_n)$

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?

For $n = 1$ we have $\hat{\theta}_{ml} = x_1$, or the value of the sample.

4. Is the maximum likelihood estimator biased?

A lifespan with a uniform distribution means that any lifespan is equally likely up to some value, θ. However, θ is always greater than the maximum x_i, which implies that $\hat{\theta}_{ml}$ will always underestimate the true θ. Therefore the estimator is biased.

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

$$E[\hat{\theta}_{ml}] = E[max(x_1, ..., x_n)] = \theta$$

6. Is the maximum likelihood estimator consistent?

For very large n the MLE would approach θ, so yes it is consistent.

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 $experienceSquare$)

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'psych'

## The following object is masked from 'package:car':
##
##     logit

## Loading required package: AER

## Loading required package: survival
```

```
df1 <- read.csv('WageData2.csv')
```

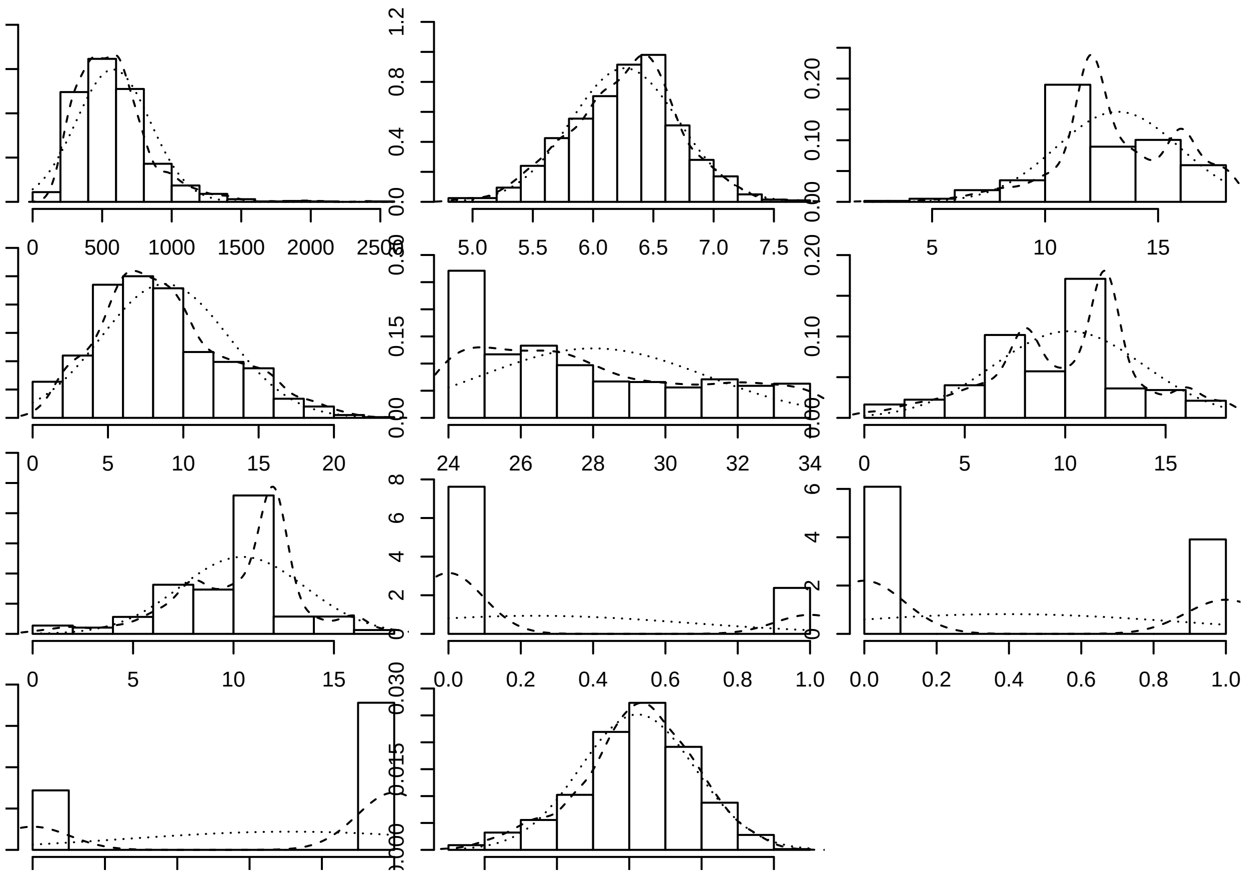
```r
str(df1)
```

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

```r
summary(df1)
```

```
##        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
```

```r
mhist <- df1[,c('wage','logWage','education','experience','age',
                'dad_education','mom_education','raceColor','rural',
                'city','IQscore')]
par(mar=c(0.5,0.5,0.5,0.5))
multi.hist(mhist, main='')
```



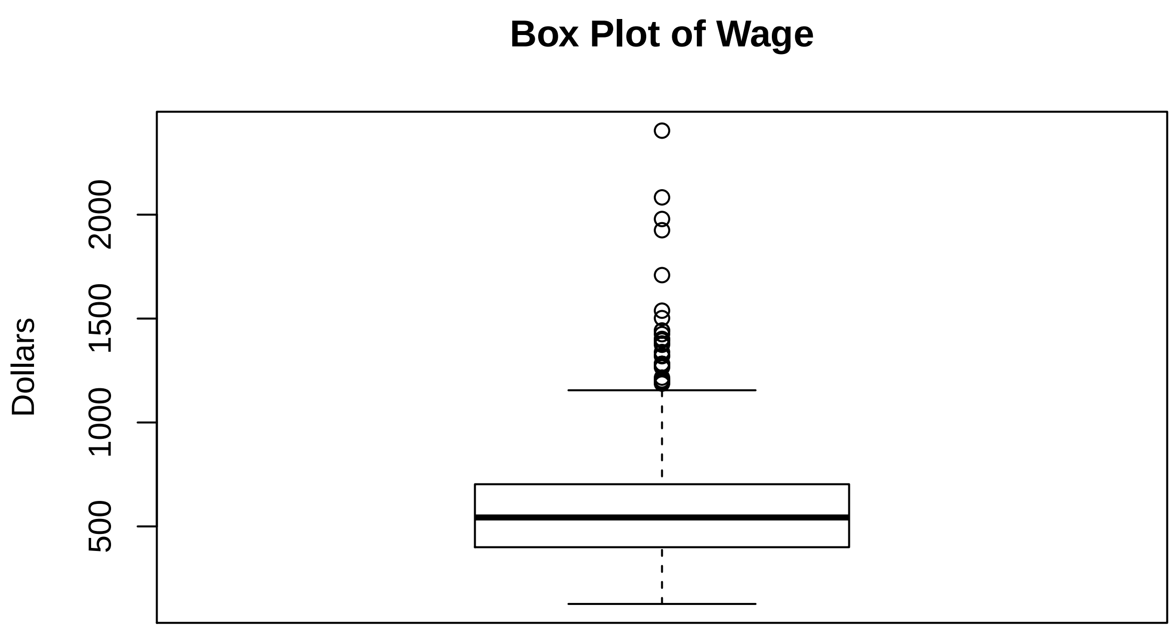
Examine the *wage* variable

```r
summary(df1$wage)
```

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

```r
hist(df1$wage, breaks=50, main='Histogram of wage', xlab='wage')
```


```
boxplot(df1$wage, main='Box Plot of Wage', ylab='Dollars')
```



Wage is right-skewed with a long tail. This will cause issues without transformation or perhaps using the *logWage* variable instead.

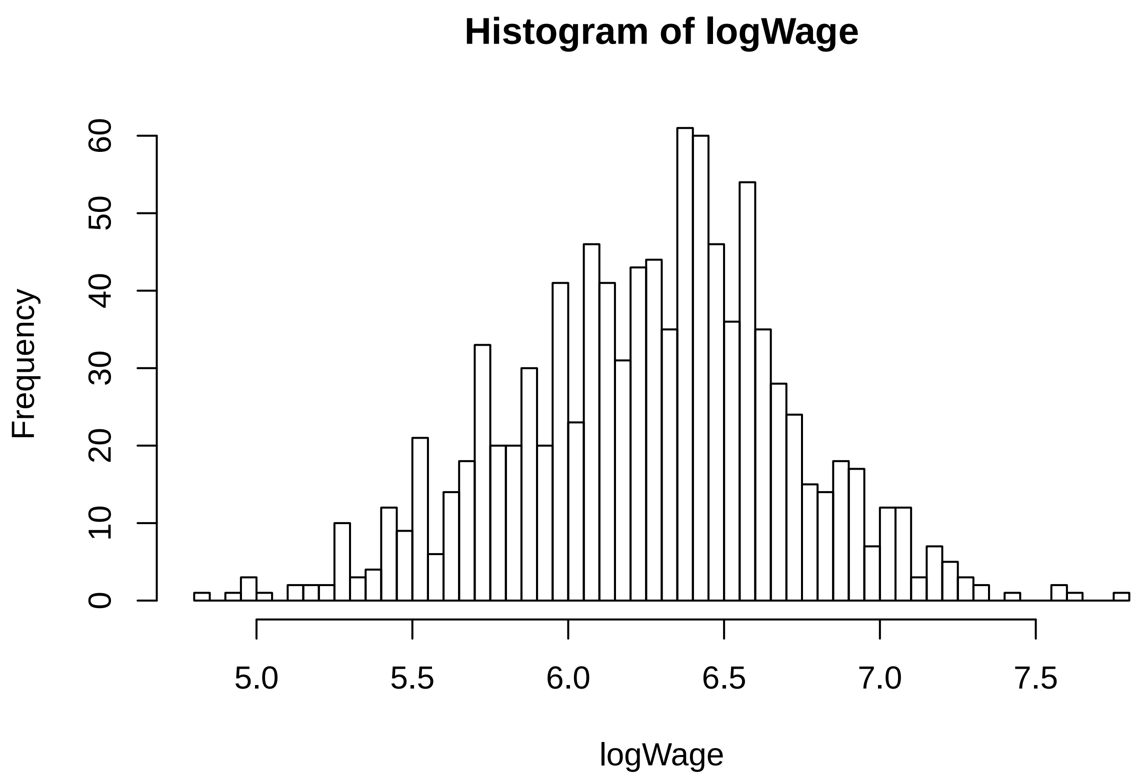
Example the *logWage* variable.

```
summary(df1$logWage)
```

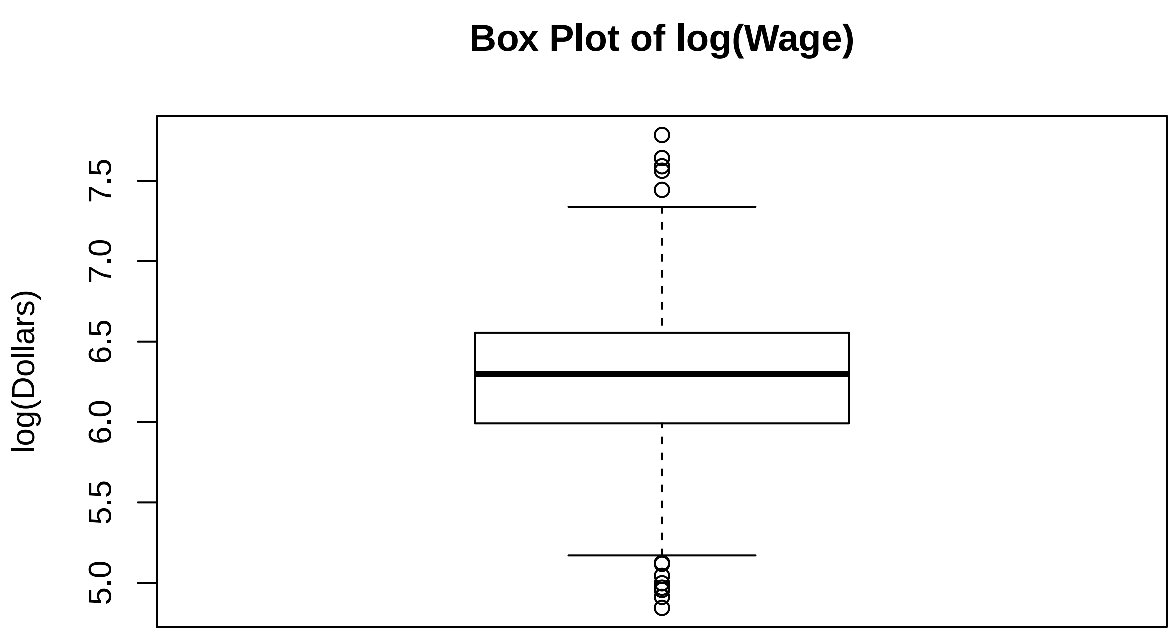
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   4.844   5.991   6.297   6.263   6.555   7.785
```

```r
hist(df1$logWage, breaks=50, main='Histogram of logWage', xlab='logWage')
```




```r
boxplot(df1$logWage, main='Box Plot of log(Wage)', ylab='log(Dollars)')
```



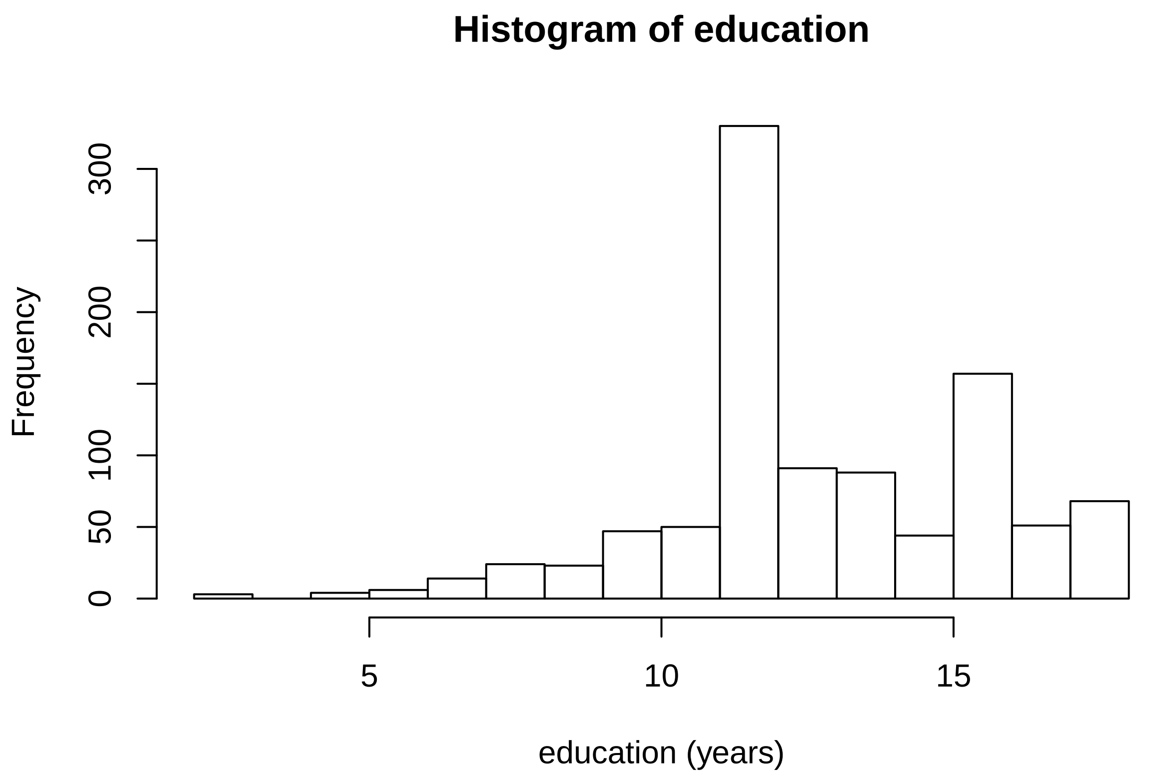
The *logWage* variable appears to correct most of the issues with the *wage* variable.

Examine the *education* variable

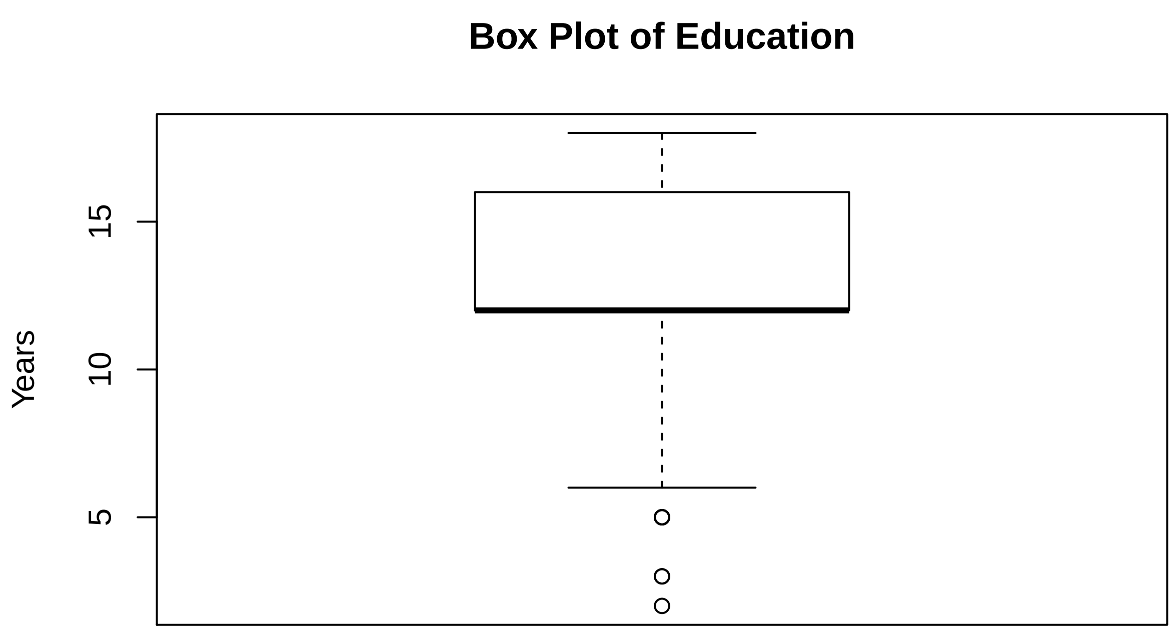
```r
summary(df1$education)
```

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

```r
hist(df1$education, breaks=18, main='Histogram of education', xlab='education (years)')
```




```r
boxplot(df1$education, main='Box Plot of Education', ylab='Years')
```



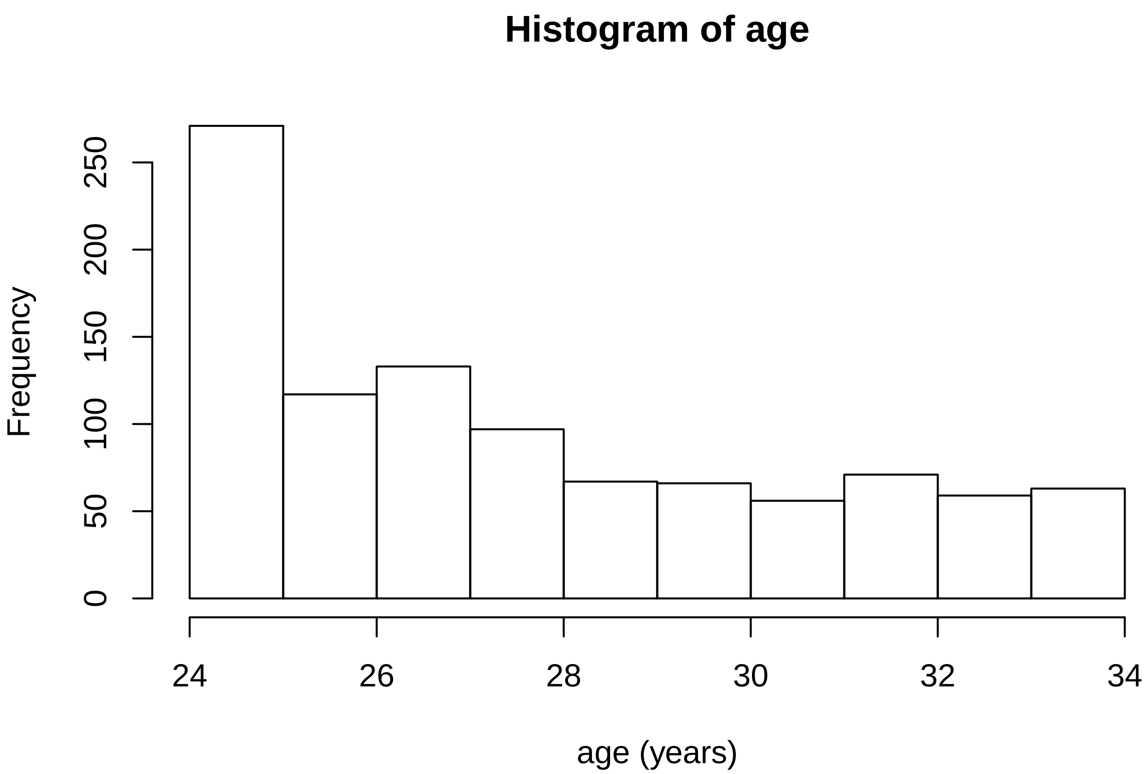
The *education* variable appears to be left-skewed with a long tail on the left. It also has a very strong peak at 12 years, corresponding to high school completion. There is a second, smaller peak at 16 years, corresponding to completing undergraduate studies.

Examine the *age* variable.

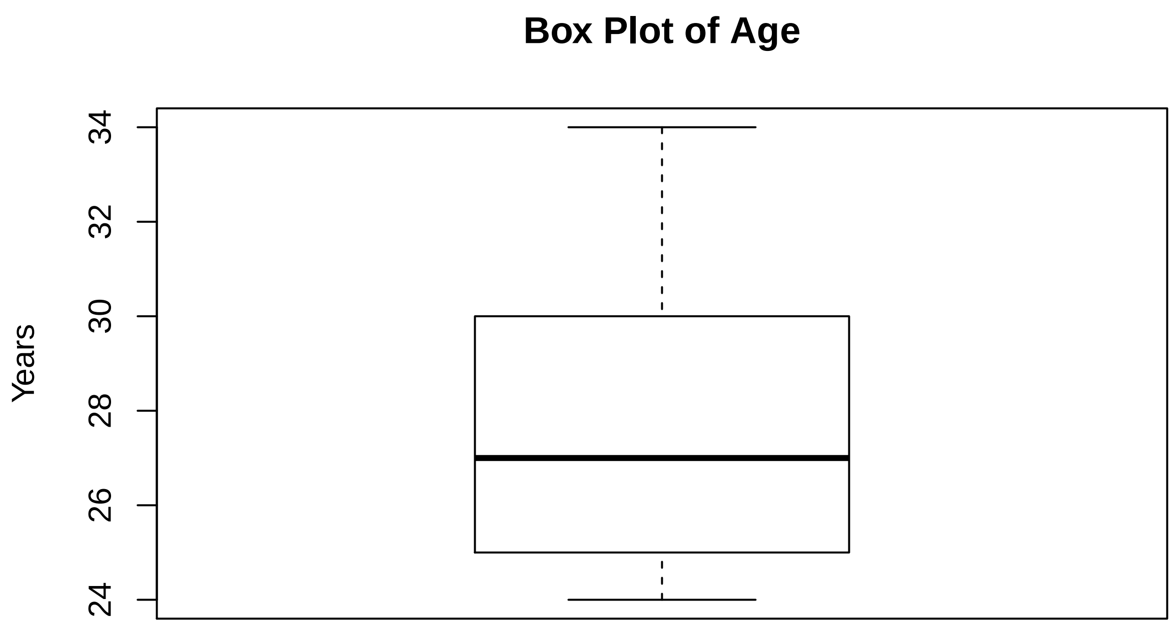
```
summary(df1$age)
```

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

```
hist(df1$age, breaks=10, main='Histogram of age', xlab='age (years)')
```




```
boxplot(df1$age, main='Box Plot of Age', ylab='Years')
```



The distribution of the *age* samples are left-skewed with a strong peak at 24 years. Therefore there will be some weighting of this analysis towards recent college graduates or high school graduates with 6 years of experience.

Examine the *raceColor* indicator variable.

```
summary(df1$raceColor)
```

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

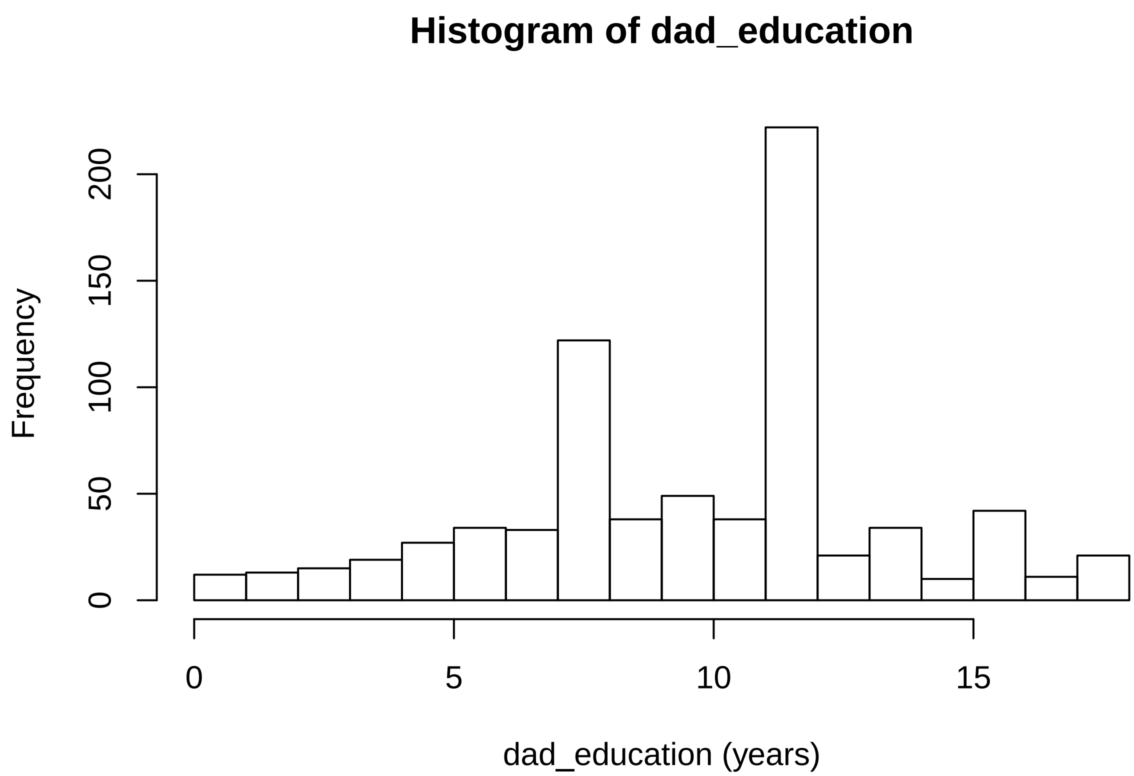
The *raceColor* variable is an indicator variable with a mean of 0.238, indicating nearly 25% non-caucasion samples in the data set.

Examine the *dad_eductation* variable

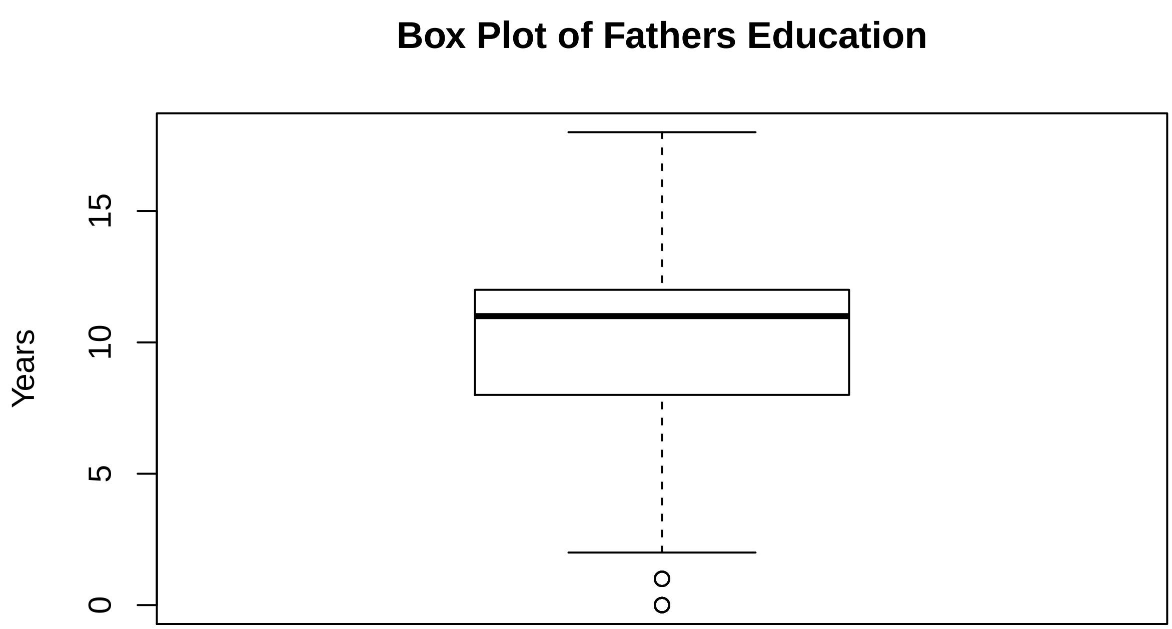
```
summary(df1$dad_education)
```

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

```
hist(df1$dad_education, breaks=20, main='Histogram of dad_education', xlab='dad_education (years)')
```




```
boxplot(df1$dad_education, main='Box Plot of Fathers Education', ylab='Years')
```



The *dad_education* variable shows a somewhat symmetrical distribution except for two strong peaks, one at 8 years and one at 12 years. This reflects the relative generation and time period of the data set when eduction beyond high school was rare and it was common to leave school after 8th grade. Nearly 24% of the samples of this variable are NA.

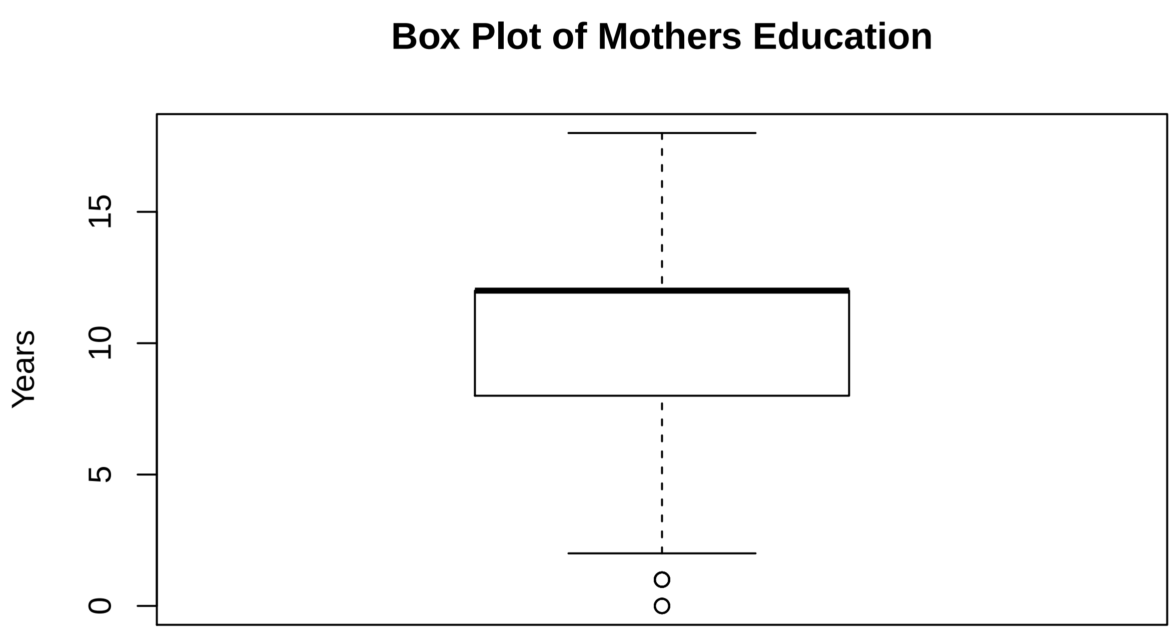
Examine *mom_education* variable

```r
summary(df1$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
hist(df1$mom_education, breaks=20, main='Histogram of mom_education', xlab='mom_education (years)')
```


```r
boxplot(df1$mom_education, main='Box Plot of Mothers Education', ylab='Years')
```



The *mom_education* variable is similar to the *dad_education* variable except the 8 year peak is much less pronounced. There are 12.8% NA's in variable samples. The box plot shows that there are definite issues with the distribution as the mean is also the first quartile.

Examine the *rural*, *city* indicator variables

```r
summary(df1$rural)
```

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

```r
summary(df1$city)
```

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

The *rural* vs. *city* percentages are 39.1% and 71.2% respectively.

Examine the *z1* and *z2* indicator variables

```r
summary(df1$z1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.44    1.00    1.00
```

```r
summary(df1$z2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   0.686   1.000   1.000
```

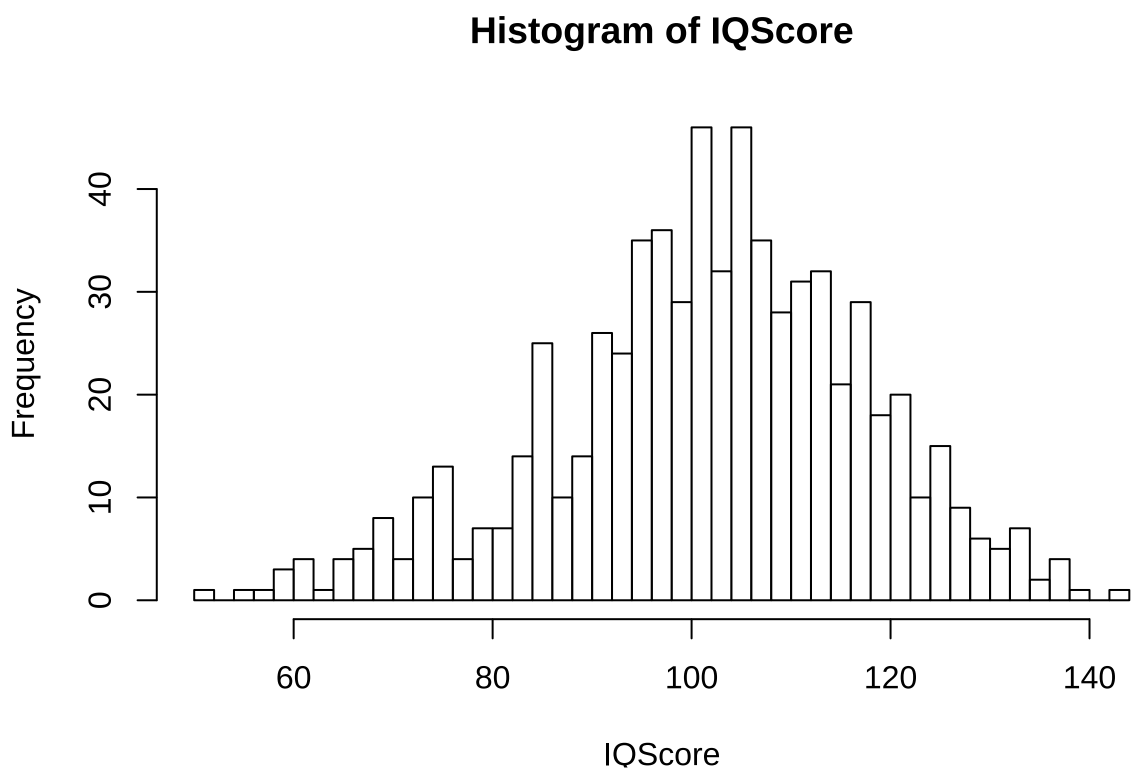
The *z1* and *z2* variables are instrument variable candidates.

Examine *IQScore* variable

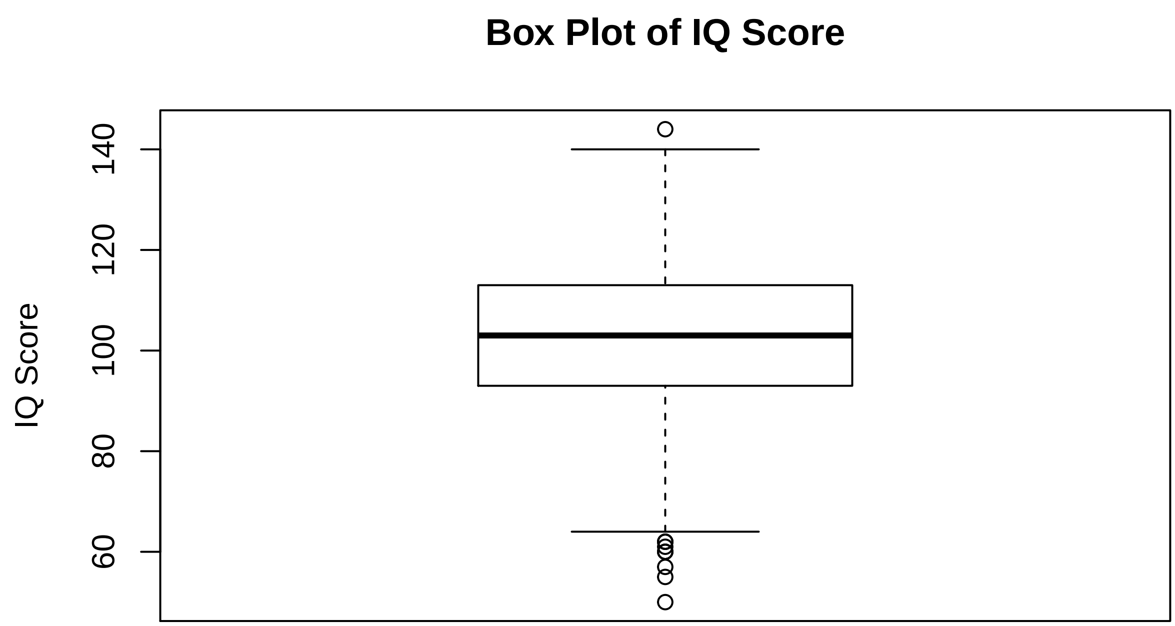
```r
summary(df1$IQscore)
```

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

```r
hist(df1$IQscore, breaks=50, main='Histogram of IQScore', xlab='IQScore')
```




```
boxplot(df1$IQscore, main='Box Plot of IQ Score', ylab='IQ Score')
```



The *IQScore* variable appears to be symmetric near the mean.

Create variables for the natural log of wage and the square of experience.

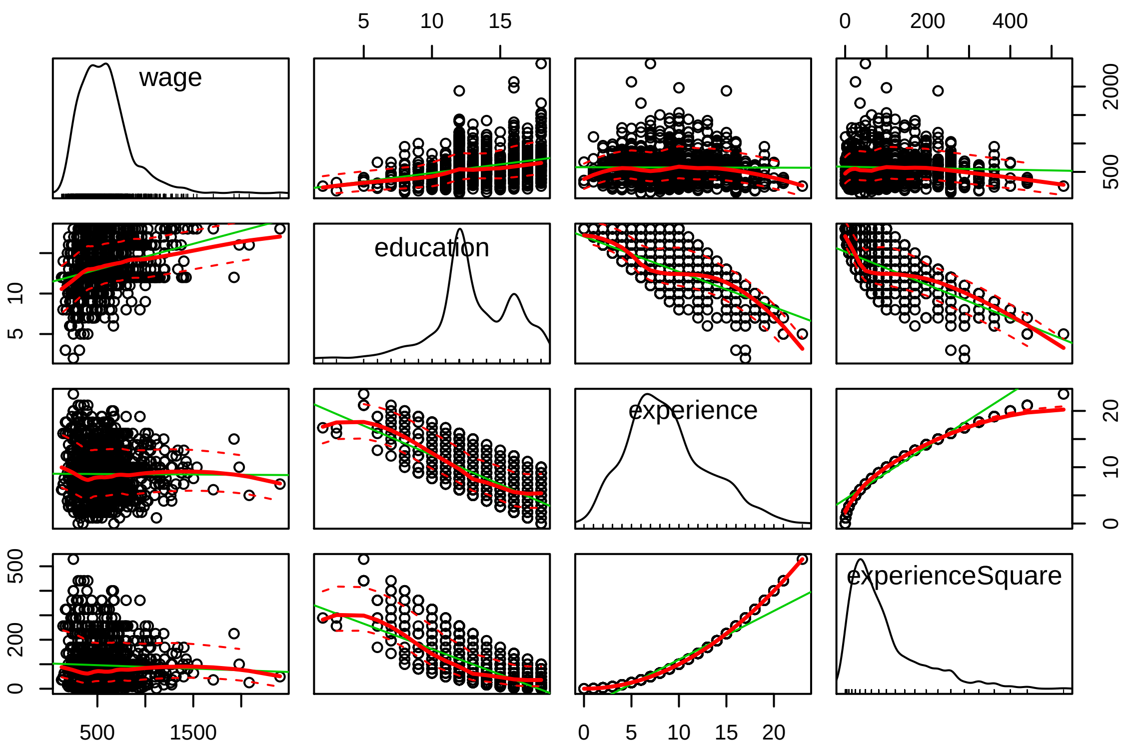
```
df1$lnWage <- log(df1$wage)
df1$experienceSquare <- df1$experience*df1$experience
```

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.

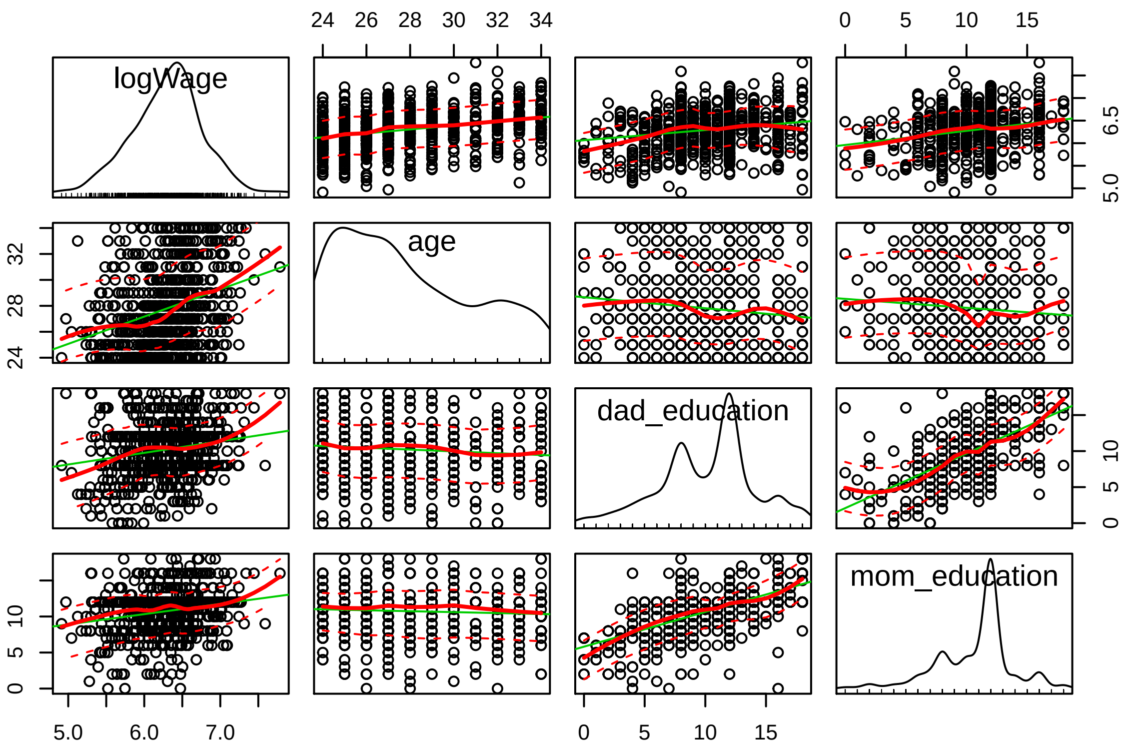
Examine *wage*, *education*, *experience*, *experienceSquare* in a scatterplot matrix

```
scatterplotMatrix(~ wage + education + experience + experienceSquare, data=df1)
```



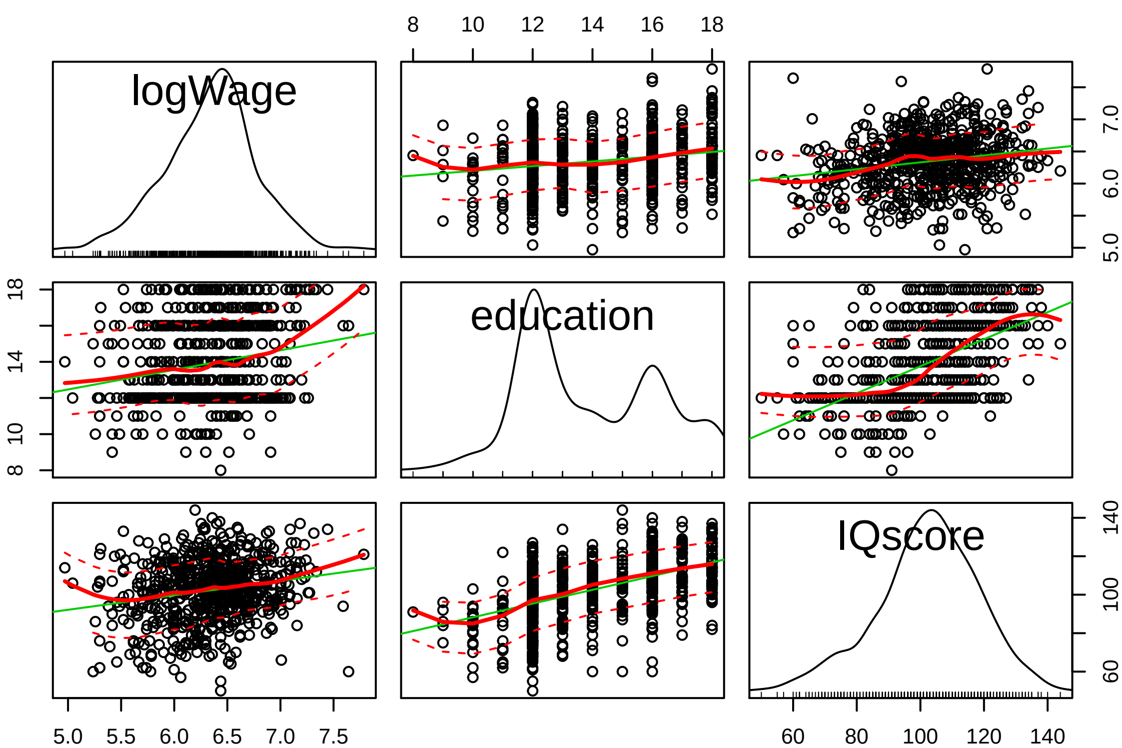
Examing *logWage*, *age*, *dad_education*, *mom_education*

```
scatterplotMatrix(~ logWage + age + dad_education + mom_education, data=df1)
```

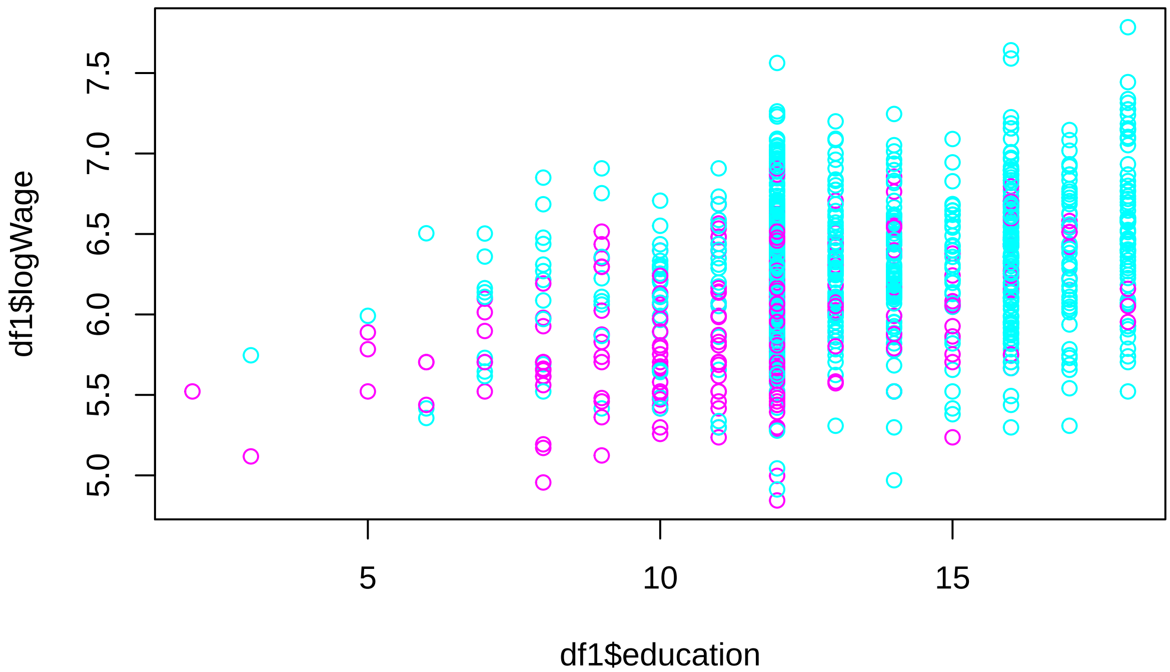


Examine *logWage*, *education*, *experience*, *raceColor*

```r
scatterplotMatrix(~ logWage + education + IQscore, data=df1)
```




```r
plot(df1$education, df1$logWage, col=ifelse(df1$raceColor==0,'cyan','magenta'))
```



Question 4.3

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

```
model4.3 <- lm(logWage ~ education + experience + age + raceColor, data=df1)
```

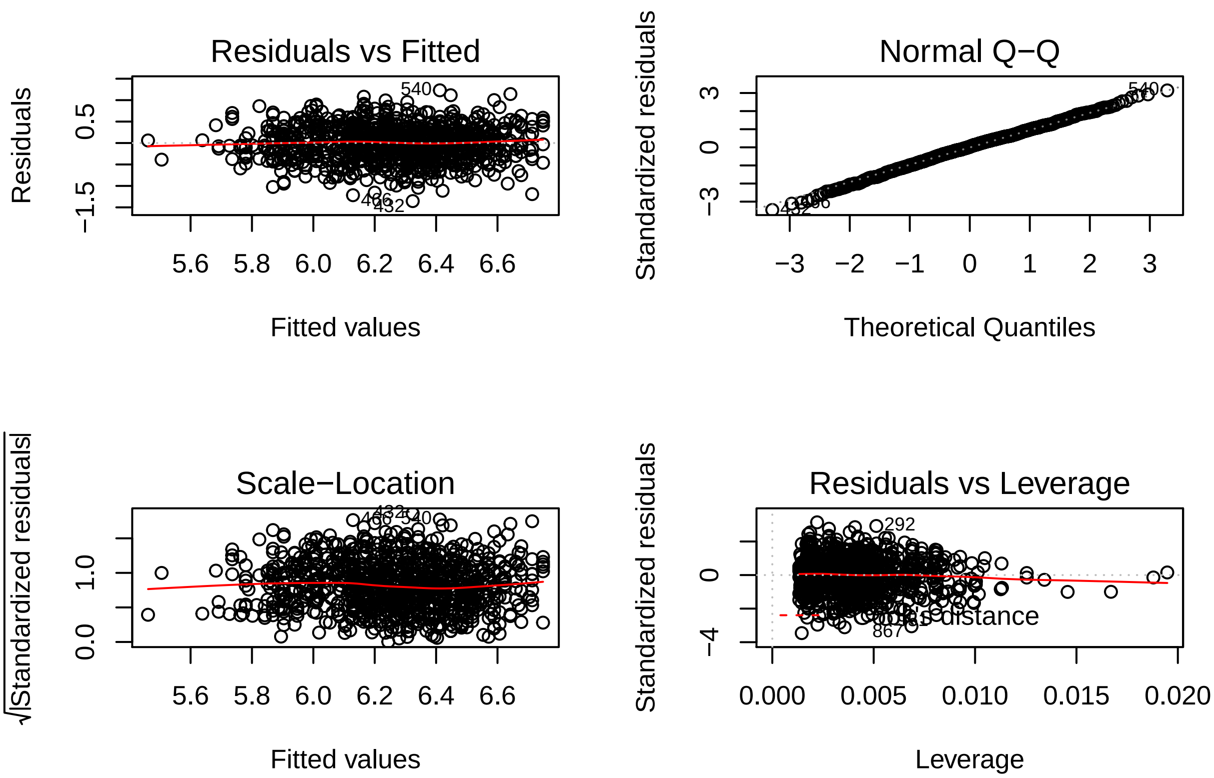
1. Report all the estimated coefficients, their standard errors, t-statistics, F-statistic of the regression, R^2, $adjusted\ R^2$, and degrees of freedom.

```
summary(model4.3)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + age + raceColor,
##     data = df1)
##
## 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
```

```r
par(mfrow = c(2,2))
plot(model4.3, sub.caption="Model Diagnostic Plots")
```



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

There are 996 degrees of freedom in the regression output, which is the size of the data set less the number of estimated parameters of the model and the intercept: $DF = 1000 - 3 - 1 = 996$

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).

The age variable is linearly related to one of the other variables which is why the coefficient for age is NA in the model summary. The model actually only estimates the coefficients for eduction, experience, raceColor and the intercept. To resolve this particular issue we drop the variable from the model, which the lm() function has done for us.

The raceColor coefficient is also a surprise at how large it is, coming in as a 26% decrease in wages, controlling for education and experience. The other coefficients are much less at 8% increase per year of eduction controlling for raceColor and experience, and 3.5% increase in wages controlling for education and raceColor. To understand the size of the coefficient for raceColor I would first analyze its contribution to the explanation of the variance. I would also investigate the interaction of race with education and race with experience to gain more insight how those factors may correlate to each other.

4. Interpret the coefficient estimate associated with education

The coefficient associated with education results in a 7.9% increase in wages controlling for experience and raceColor, and is highly statistically significant.

5. Interpret the coefficient estimate associated with experience

The coefficient associated with experience results in a 3.5% increase in wages controlling for education and raceColor, and is highly statistically significant.

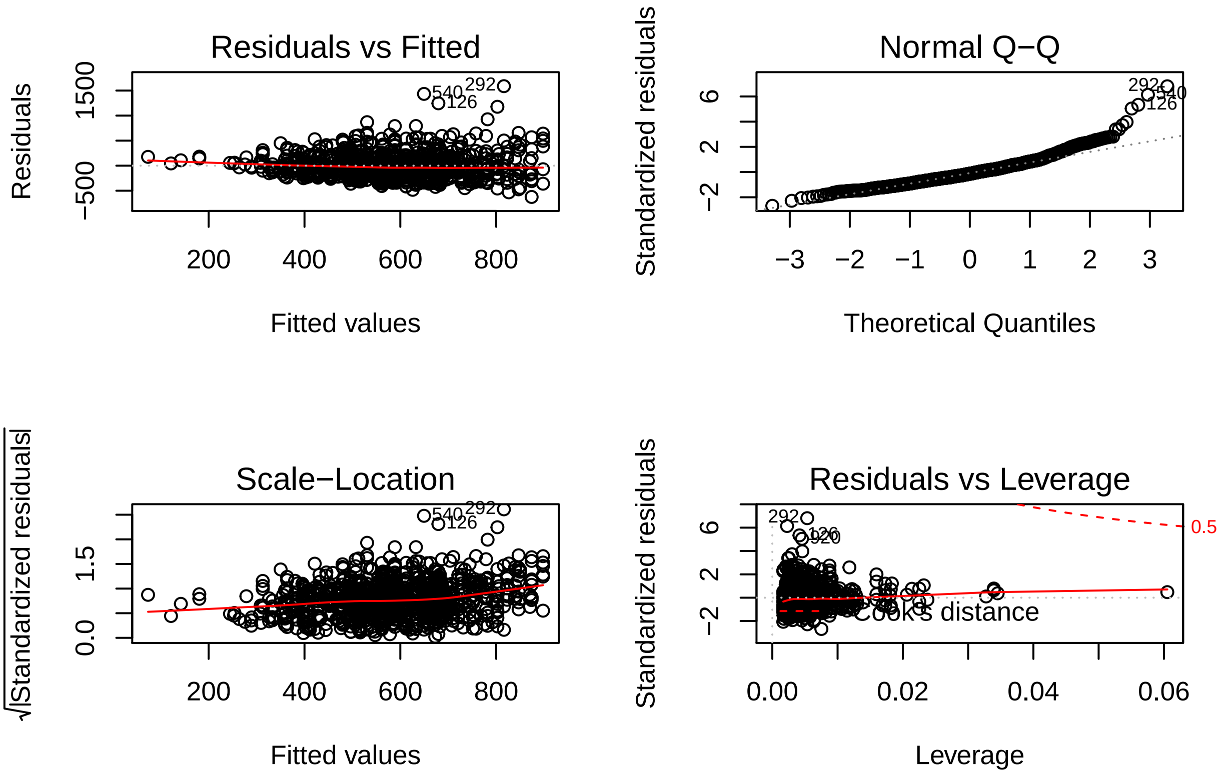
Question 4.4

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

```r
model4.4 <- lm(wage ~ education + experience + experienceSquare + raceColor, data=df1)
summary(model4.4)
```

```
##
## Call:
## lm(formula = wage ~ education + experience + experienceSquare +
##     raceColor, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -624.02 -150.34  -30.92  113.71 1587.82
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -351.2296    72.4556  -4.848 1.45e-06 ***
## education          47.5871     3.8062  12.503  < 2e-16 ***
## experience         56.4494     6.9658   8.104 1.55e-15 ***
## experienceSquare   -1.7206     0.3298  -5.217 2.21e-07 ***
## raceColor        -132.8207    18.1804  -7.306 5.65e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 233.8 on 995 degrees of freedom
## Multiple R-squared:  0.2337, Adjusted R-squared:  0.2306
## F-statistic: 75.85 on 4 and 995 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(model4.4, sub.caption="Model Diagnostic Plots")
```



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

```
plot(x=NULL, y=NULL, xlim=c(0,30),ylim=c(0,1500), ylab='$', xlab='Years', main='Plot of Experience Coef
abline(a = coef(model4.4)[1], b = coef(model4.4)[3], lwd = 2, col = "red")
abline(v = 10, lwd = 2, col = "blue")
```

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

```
coef(model4.4)[1]+10*coef(model4.4)[3]
```

```
## (Intercept)
##    213.2644
```

$213.26 increase in wage for 10 years of experience

Question 4.5

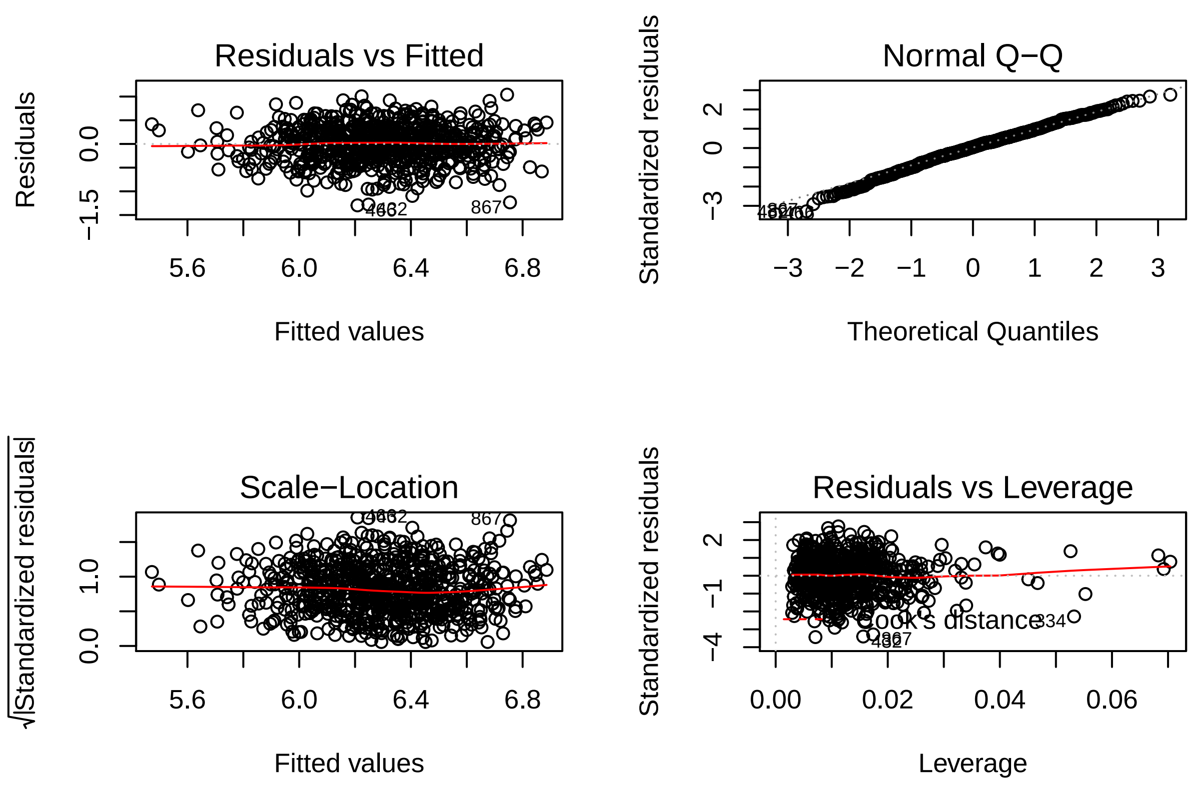
Regress *logWage* on *education*, *experience*, *experienceSquare*, *raceColor*, *dad_education*, *mom_education*, *rural*, *city*.

```
model4.5 <- lm(logWage ~ education + experience + experienceSquare + raceColor +
                 dad_education + mom_education + rural + city, data=df1)
summary(model4.5)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education + mom_education + rural + city,
##     data = df1)
##
## 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
```

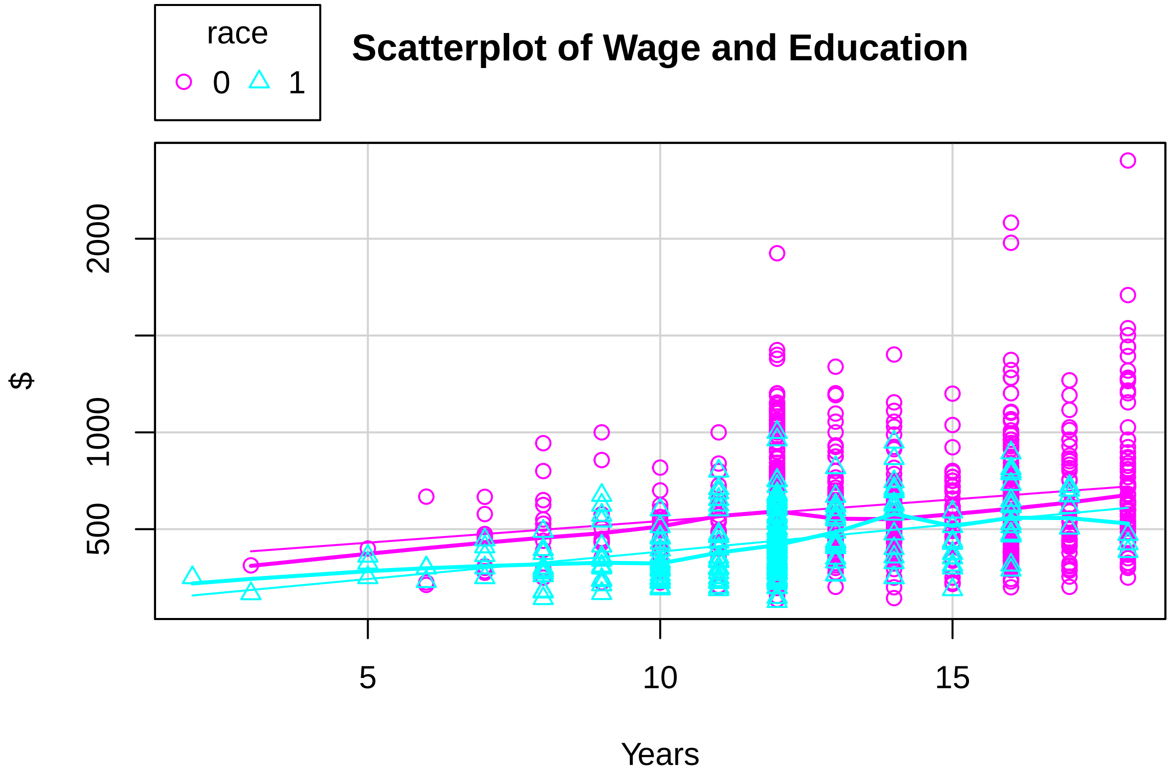
```r
par(mfrow = c(2,2))
plot(model4.5, sub.caption="Model Diagnostic Plots")
```



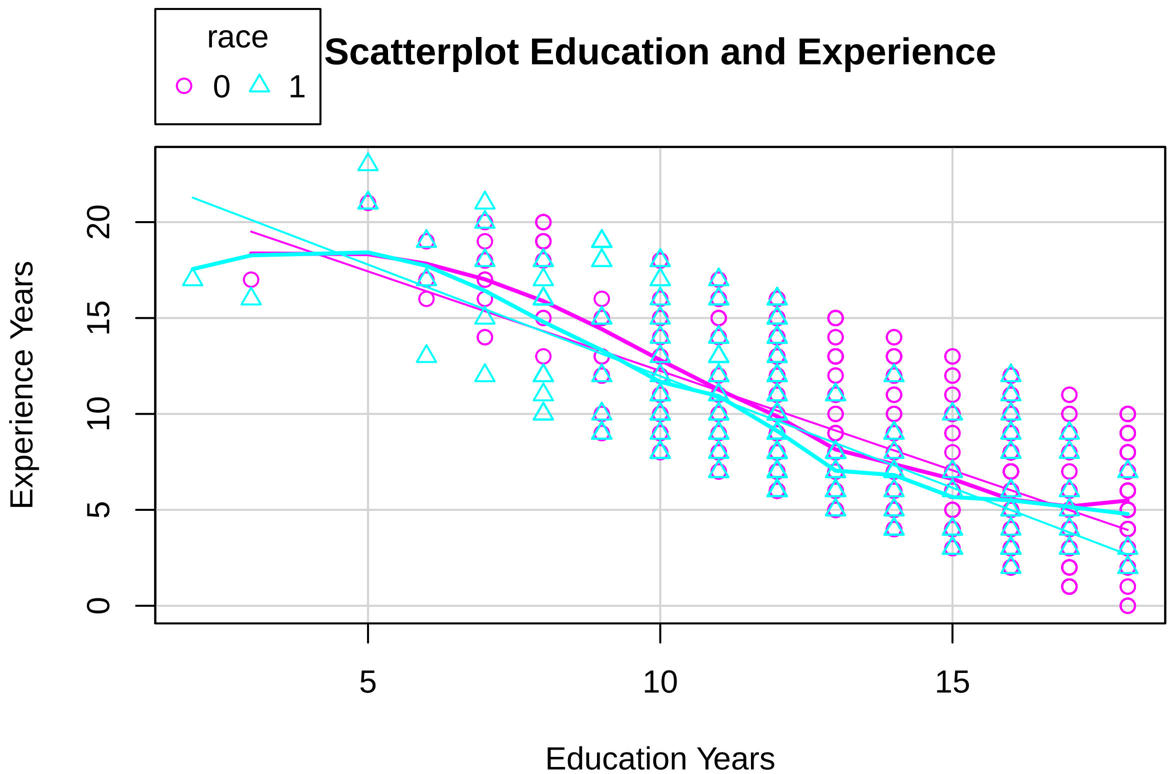
1. What are the number of observations used in this regression? Are missing values a problem? Analyze the missing values, if any, and see if there is any discernible pattern with wage, education, experience, and raceColor.

There are 1000 - 277 = 723 observations.

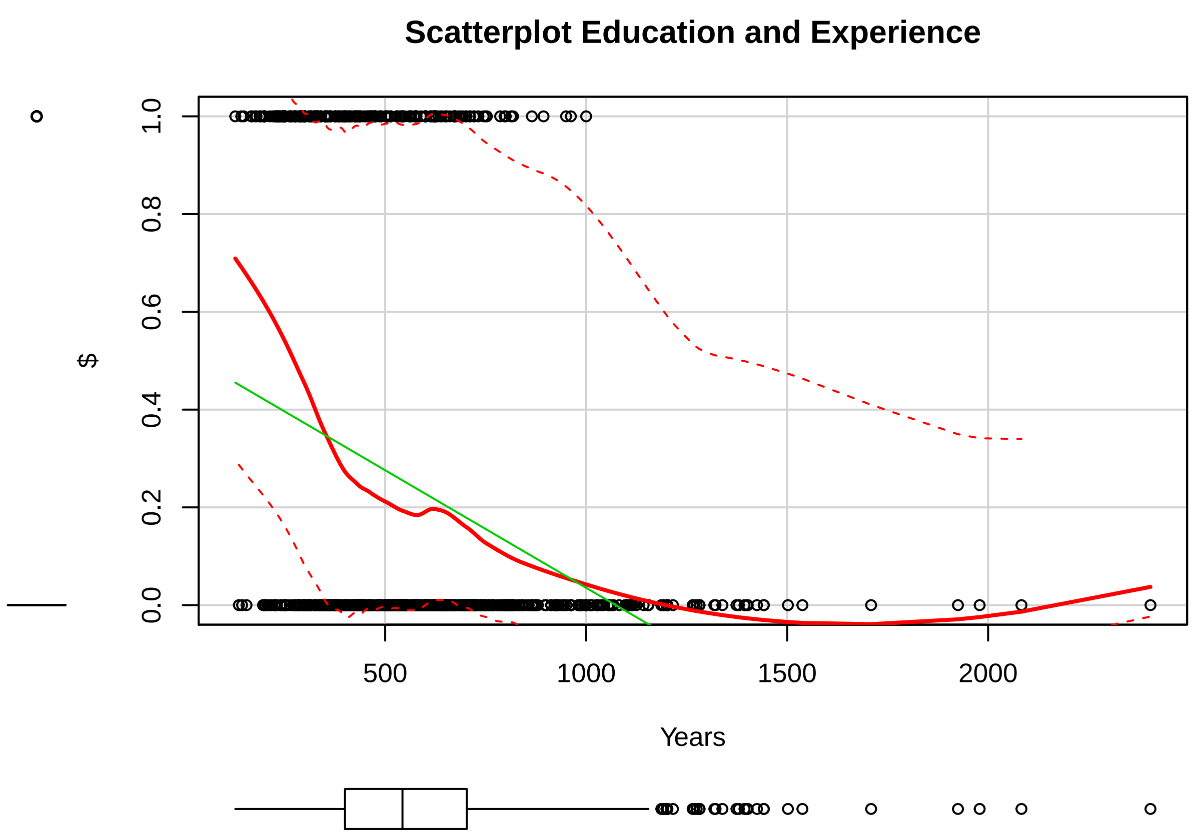
```r
scatterplot(df1$education, df1$wage, groups=df1$raceColor, by.groups=TRUE,
            xlab='Years', ylab='$', main='Scatterplot of Wage and Education',
            legend.title='race', col=c('magenta','cyan'))
```




```r
scatterplot(df1$education, df1$experience, groups=df1$raceColor, by.groups=TRUE,
            xlab='Education Years', ylab='Experience Years', main='Scatterplot Education and Experience
            legend.title='race', col=c('magenta','cyan'))
```




```
scatterplot(df1$wage, df1$raceColor, xlab='Years', ylab='$', main='Scatterplot Education and Experience
```



There is a fixed linear relationship between education and experience.

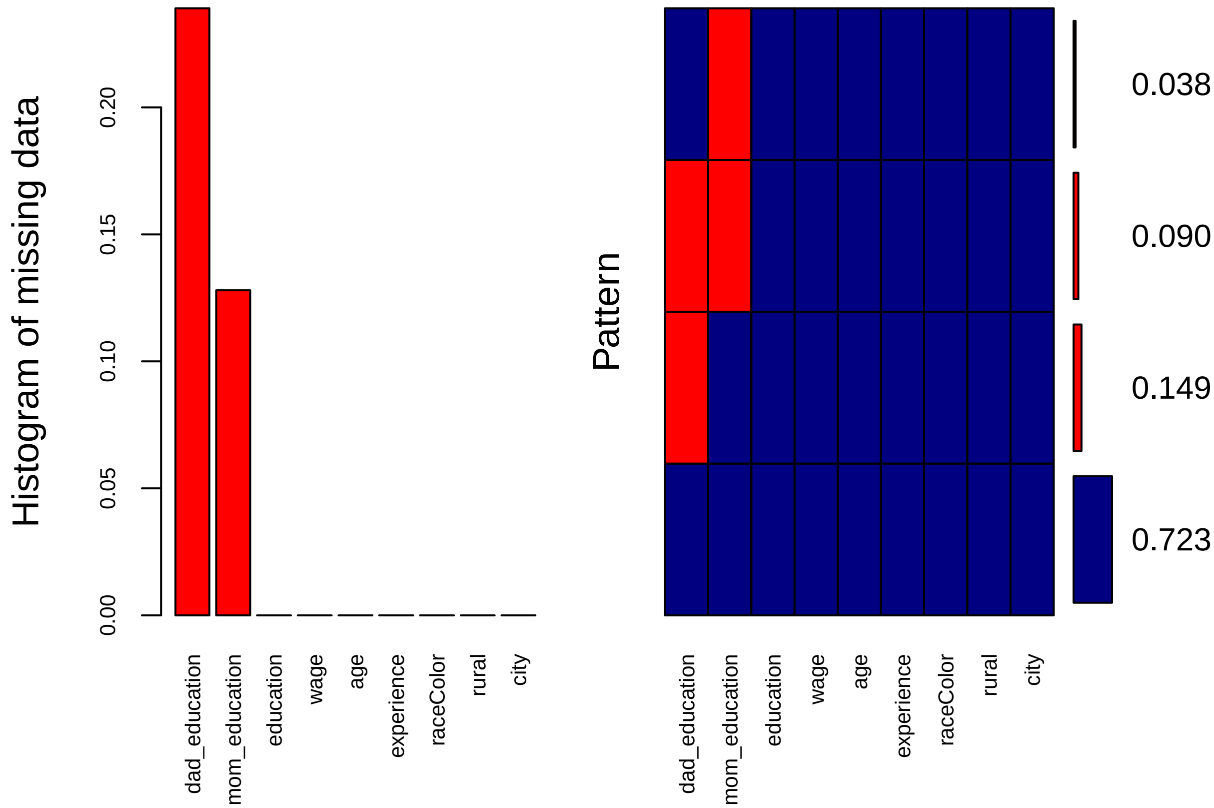
```r
library(mice)
```

```
## Loading required package: Rcpp

## mice 2.25 2015-11-09
```

```r
library(VIM)
```

```
## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```r
aggr_plot <- aggr(df1[,c('education','wage','dad_education','mom_education',
                         'age','experience','raceColor','rural','city')],
                  col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE,
                  labels=names(data), cex.axis=.7, gap=3,
                  ylab=c("Histogram of missing data","Pattern"))
```




```
##
## Variables sorted by number of missings:
##       Variable Count
##  dad_education 0.239
##  mom_education 0.128
##      education 0.000
##           wage 0.000
##            age 0.000
##     experience 0.000
##      raceColor 0.000
##          rural 0.000
##           city 0.000
```

Over 25% of the dad_education data is missing, and about 12% of the mom_education data.

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

The NA observations have value but we're talking about a large portion of the values for those 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?

One could replace the NA values with the mean of the variable but additional bias is introduced on top of what bias may already exist. Adding a large number of mean values will also reduce the variance.

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?

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

```r
summary(model4.3)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + age + raceColor,
##     data = df1)
##
## 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
```

```r
summary(model4.4)
```

```
##
## Call:
## lm(formula = wage ~ education + experience + experienceSquare +
##     raceColor, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -624.02 -150.34  -30.92  113.71 1587.82
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -351.2296    72.4556  -4.848 1.45e-06 ***
## education          47.5871     3.8062  12.503  < 2e-16 ***
## experience         56.4494     6.9658   8.104 1.55e-15 ***
## experienceSquare   -1.7206     0.3298  -5.217 2.21e-07 ***
## raceColor        -132.8207    18.1804  -7.306 5.65e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 233.8 on 995 degrees of freedom
## Multiple R-squared:  0.2337, Adjusted R-squared:  0.2306
## F-statistic: 75.85 on 4 and 995 DF,  p-value: < 2.2e-16
```

```r
summary(model4.5)
```

```
##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education + mom_education + rural + city,
##     data = df1)
##
## 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
```

I prefer the first model, from Question 4.3 because it has the highest F-statistic even though it doesn't have the highest Adjusted R^2. It is also the simpler model with the fewest parameters (most parsimonious).

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)?

The assumptions to be satisfied are $cov(z_1, u) = 0$ and that if the variable for which we want to use z_i as an indicator is x then $cov(x, z_1) \neq 0$

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?

There are several scenarios in which z_1 could be correlated with the error term, u. - Adjacent regions and policy may have an effect, especially in regions with more industry or higher paying jobs. There is more money to be spent on education in that case. There is no per capita expenditure on education variable in the data set so this would be in the error term. - Local and regional attitudes can also have an effect in the error term, such as whether the region contains a college or university town. People in these localities may have a propensity to favor emphasis on education.

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?

First let's check the relationship between z_1 zand z_2 on the outcome variable.

```
model4.6_z1 <- lm(logWage ~ z1, data=df1)
summary(model4.6_z1)
```

```
##
## Call:
## lm(formula = logWage ~ z1, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46232 -0.28225  0.03737  0.28380  1.47838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.22840    0.01885 330.465  < 2e-16 ***
## z1           0.07810    0.02841   2.749  0.00609 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.446 on 998 degrees of freedom
## Multiple R-squared:  0.007514,   Adjusted R-squared:  0.00652
## F-statistic: 7.556 on 1 and 998 DF,  p-value: 0.006089
```

There is a statistically significant relationship between z_1 and *education*.

```
model4.6_z2 <- lm(logWage ~ z2, data=df1)
summary(model4.6_z2)
```

```
##
## Call:
## lm(formula = logWage ~ z2, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47200 -0.28529  0.04079  0.29008  1.46871
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.14607    0.02487 247.146  < 2e-16 ***
## z2           0.17011    0.03002   5.666 1.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4407 on 998 degrees of freedom
## Multiple R-squared:  0.03116,    Adjusted R-squared:  0.03019
## F-statistic:  32.1 on 1 and 998 DF,  p-value: 1.915e-08
```

There is a highly statistically significant relationship between z_2 and *education*.

Performing a 2SLS regression using both z_1 and z_2

```
model4.6_iv <- ivreg(logWage ~ education + experience + experienceSquare + raceColor
                     + dad_education + mom_education + rural + city | z1 + z2, data=df1)
```

```
## Warning in ivreg.fit(X, Y, Z, weights, offset, ...): more regressors than
## instruments
```

```
robust.se(model4.6_iv)
```

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

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.24441         NA      NA       NA
## education         0.35165         NA      NA       NA
## experience        0.45926         NA      NA       NA
## experienceSquare       NA         NA      NA       NA
## raceColor              NA         NA      NA       NA
## dad_education          NA         NA      NA       NA
## mom_education          NA         NA      NA       NA
## rural                  NA         NA      NA       NA
## city                   NA         NA      NA       NA
```

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 Exploration

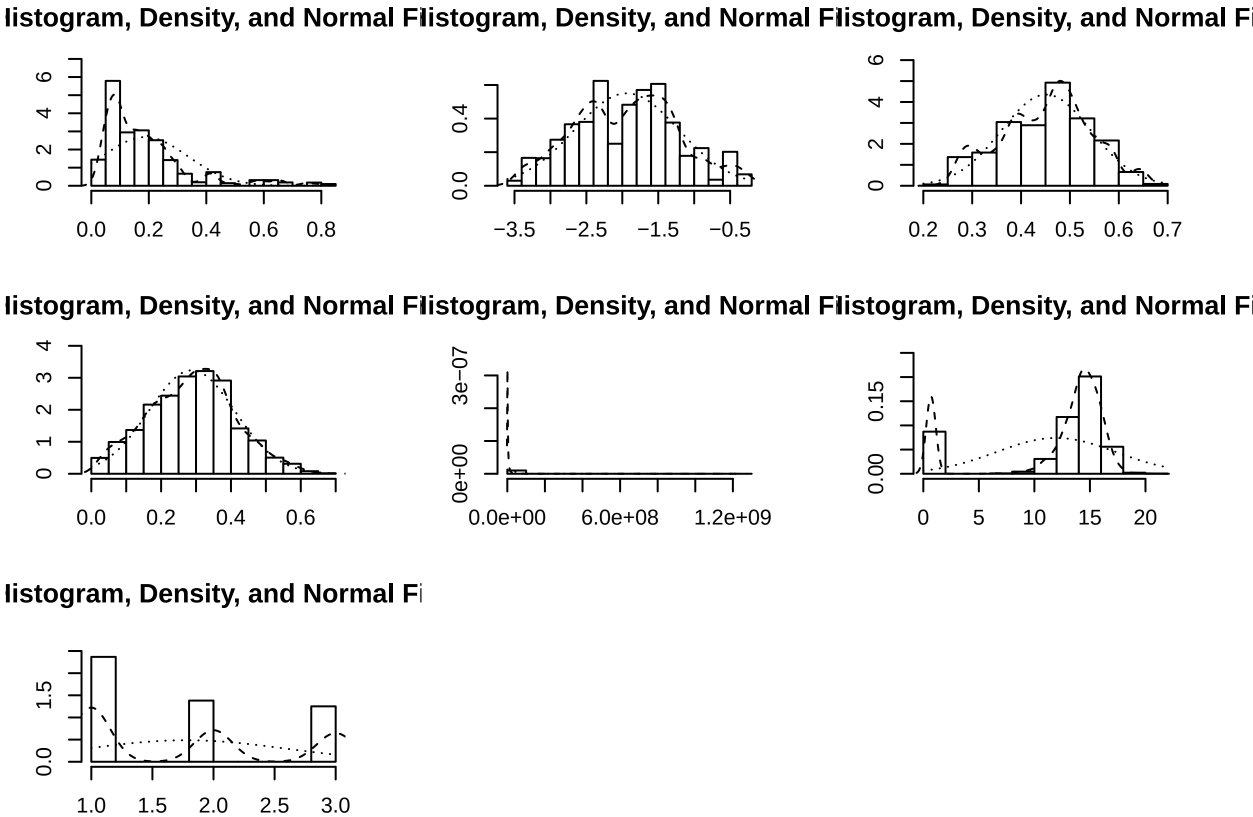
```
df2 <- read.csv('wealthy_candidates.csv')
str(df2)
```

```
## '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(df2)
```

```
##        X               region          urb               lit
##  Min.   :   1.0   Region 1:1183   Min.   :0.02835   Min.   :0.2418
##  1st Qu.: 625.2   Region 2: 690   1st Qu.:0.08387   1st Qu.:0.3846
##  Median :1249.5   Region 3: 625   Median :0.14657   Median :0.4602
##  Mean   :1249.5                   Mean   :0.18729   Mean   :0.4512
##  3rd Qu.:1873.8                   3rd Qu.:0.24319   3rd Qu.:0.5105
##  Max.   :2498.0                   Max.   :0.80234   Max.   :0.6524
##
##    voteshare        absolute_wealth
##  Min.   :0.006037   Min.   :2.000e+00
##  1st Qu.:0.199620   1st Qu.:1.875e+05
##  Median :0.293398   Median :1.337e+06
##  Mean   :0.287860   Mean   :5.034e+06
##  3rd Qu.:0.367978   3rd Qu.:4.092e+06
##  Max.   :0.693324   Max.   :1.216e+09
##                     NA's   :1
```

```
df2$logWealth <- log(df2$absolute_wealth)
df2$logUrb<- log(df2$urb)
par(mar=c(3,3,3,3))
mhist <- df2[,c('urb','logUrb','lit','voteshare','absolute_wealth', 'logWealth')]
mhist$region <- as.numeric(df2$region)
multi.hist(mhist)
```



We can see that *urb* and *absolute_wealth* have issues with distribution. Creating a new variable, *logUrb*, as the log(urb) does help with the distribution. However, *absolute_wealth* doesn't get as much help from this treatment.

If we examing the *absolute_wealth* variable we can see there are a large number of 2.0e+00 values, which is also shown as the minimum value in the summary.

```
sum(na.omit(df2$absolute_wealth==2.0))
```

```
## [1] 435
```

```
sum(na.omit(df2$absolute_wealth > 2.0))
```

```
## [1] 2062
```

```
sum(na.omit(df2$absolute_wealth > 200.0))
```

```
## [1] 2062
```

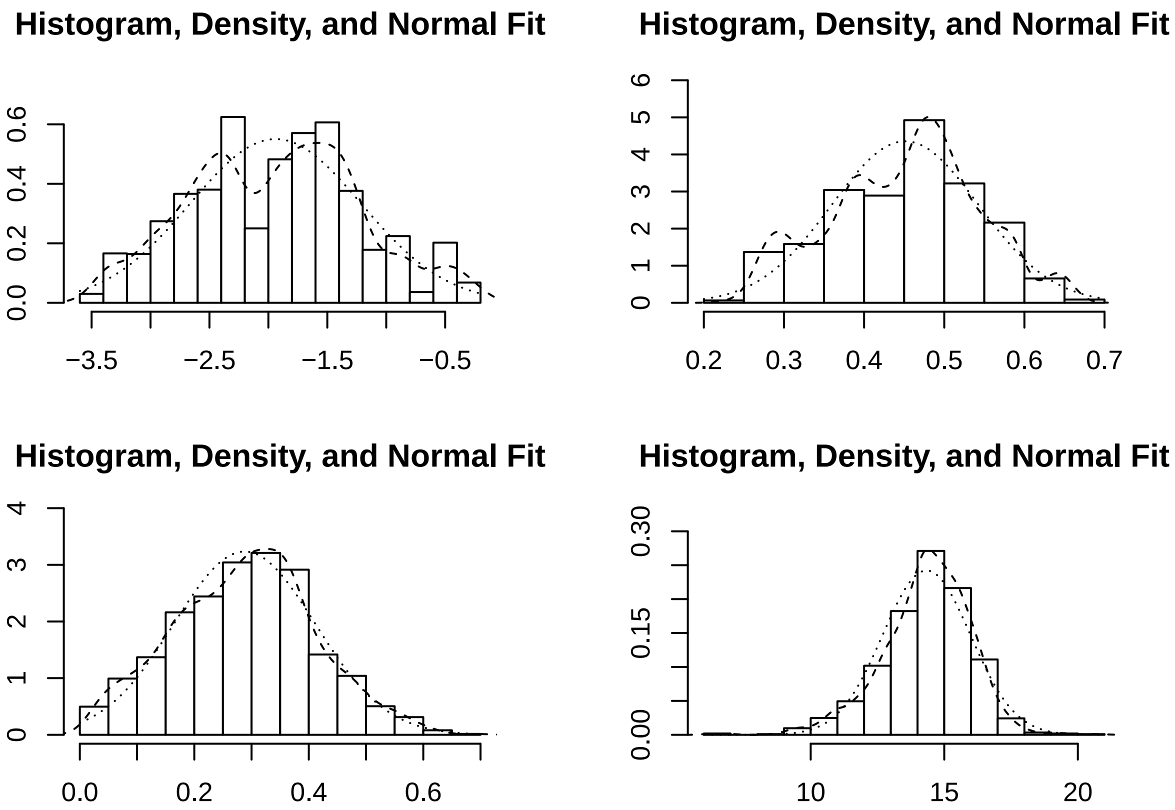
There are 435 entries with a value of 2.0 and there are 2062 entries with values greater than 2.0 and also greater than 200.0. The count starts dropping slightly around 2000.0. There are also 162 NA values.

We don't have any information about how this value for *absolute_wealth* came to be so our best course of action is to set it to NA.

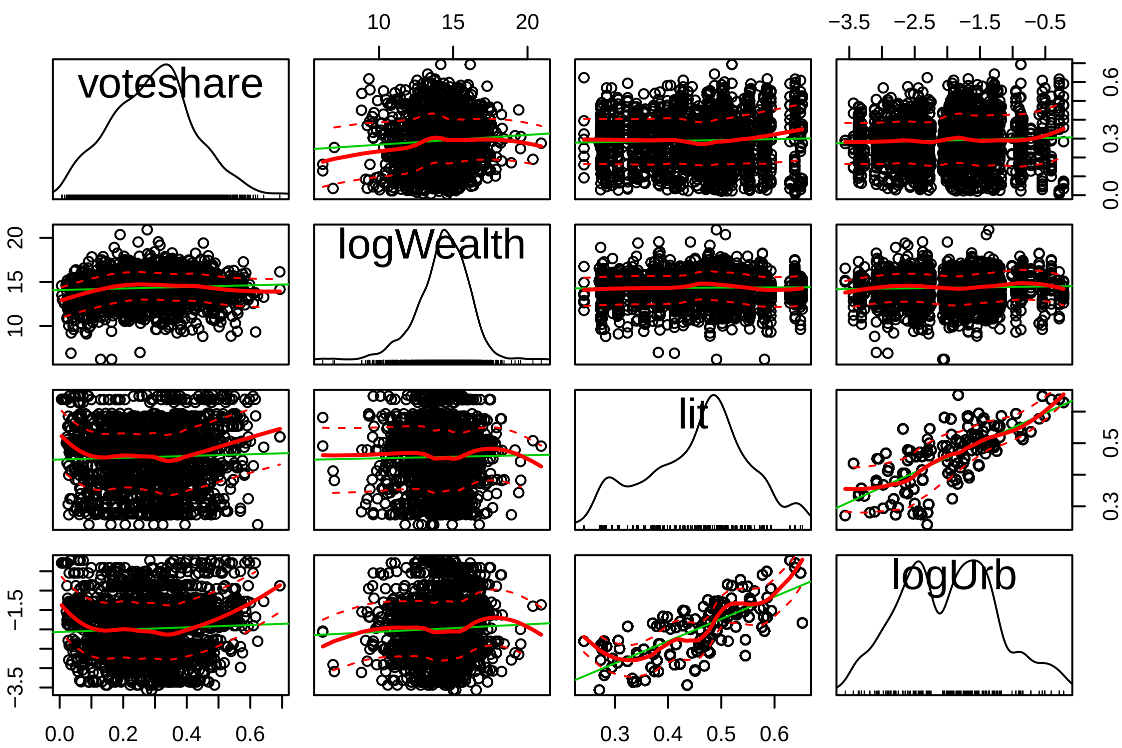
```r
df2$wealth<-df2$absolute_wealth
df2$wealth[df2$wealth==2.0] <- NA
df2$logWealth <- log(df2$wealth)
summary(df2)
```

```
##        X                region          urb                lit
##  Min.   :   1.0   Region 1:1183   Min.   :0.02835   Min.   :0.2418
##  1st Qu.: 625.2   Region 2: 690   1st Qu.:0.08387   1st Qu.:0.3846
##  Median :1249.5   Region 3: 625   Median :0.14657   Median :0.4602
##  Mean   :1249.5                   Mean   :0.18729   Mean   :0.4512
##  3rd Qu.:1873.8                   3rd Qu.:0.24319   3rd Qu.:0.5105
##  Max.   :2498.0                   Max.   :0.80234   Max.   :0.6524
##
##    voteshare        absolute_wealth       logWealth          logUrb
##  Min.   :0.006037   Min.   :2.000e+00   Min.   : 6.217   Min.   :-3.5632
##  1st Qu.:0.199620   1st Qu.:1.875e+05   1st Qu.:13.444   1st Qu.:-2.4785
##  Median :0.293398   Median :1.337e+06   Median :14.442   Median :-1.9202
##  Mean   :0.287860   Mean   :5.034e+06   Mean   :14.338   Mean   :-1.9387
##  3rd Qu.:0.367978   3rd Qu.:4.092e+06   3rd Qu.:15.456   3rd Qu.:-1.4139
##  Max.   :0.693324   Max.   :1.216e+09   Max.   :20.919   Max.   :-0.2202
##                     NA's   :1           NA's   :436
##      wealth
##  Min.   :5.010e+02
##  1st Qu.:6.900e+05
##  Median :1.871e+06
##  Mean   :6.096e+06
##  3rd Qu.:5.157e+06
##  Max.   :1.216e+09
##  NA's   :436
```

```r
par(mar=c(3,3,3,3))
mhist <- df2[,c('logUrb','lit','voteshare', 'logWealth')]
multi.hist(mhist)
```




```r
scatterplotMatrix(~voteshare + logWealth + lit + logUrb, data=df2)
```



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?

```
model5.1 <- lm(voteshare ~ logWealth, data=df2)
coeftest(model5.1, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.2161814  0.0266030  8.1262 7.554e-16 ***
## logWealth   0.0051640  0.0018029  2.8642  0.004223 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a linear relationship between *logWealth* and *voteshare*, statistically significant at the .002 level.

The most parsimonious and direct model is the simple OLS model that expresses the relationship between vote share and wealth.

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?

We can use a quadratic of wealth to measure marginal effects.

```
model5.1.2 <- lm(voteshare ~ logWealth + I(logWealth**2), data=df2)
coeftest(model5.1.2, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.0137756  0.1224262 -0.1125  0.91042
## logWealth        0.0387796  0.0170734  2.2714  0.02323 *
## I(logWealth^2) -0.0012100  0.0005946 -2.0350  0.04198 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results indicate a stastistically significant relationship at the .04 level for $log(wealth)^2$. The coefficients suggest a diminishing return to voteshare and the point at which the return to voteshare becomes 0 is

```
abs(coef(model5.1.2)[2]/2*coef(model5.1.2)[3])
```

```
##    logWealth
## 2.346147e-05
```

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.

Using the region variable we can convert it to a set of dummy variables: *region1*, *region2*, *region3*

```r
df2$region1 <- ifelse(df2$region=="Region 1",1,0)
df2$region2 <- ifelse(df2$region=="Region 2",1,0)
df2$region3 <- ifelse(df2$region=="Region 3",1,0)
model5.1.3 <- lm(voteshare ~ logWealth + region2 + region3 + region1, data=df2)
coeftest(model5.1.3, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.0875629  0.0302076  2.8987  0.003787 **
## logWealth   0.0120376  0.0019759  6.0921 1.327e-09 ***
## region2     0.0405620  0.0065052  6.2353 5.455e-10 ***
## region3     0.0608416  0.0073576  8.2692 2.390e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the coefficients of the model are highly significant.

Need to interpret what logWealth as a parameter means in this case.

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?

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?

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.

Exploratory Data Analysis

```r
load('retailSales.Rdata')
str(retailSales)
```

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