

# 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  $X$ ,  $E(Y|X)$ .

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

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

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

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:

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

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

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

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

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

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.

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

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

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

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:

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

Given  $y$  we know that  $x$  must be on the interval  $[y, 1]$ . Therefore we have:

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

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

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

Therefore  $E(X|Y=\frac{1}{2}) = \frac{3}{4}$

## Question 2: Investing

Suppose that you are planning an investment in three different companies. The payoff per unit you invest in each company is represented by a random variable.  $A$  represents the payoff per unit invested in the first company,  $B$  in the second, and  $C$  in the third.  $A$ ,  $B$ , and  $C$  are independent of each other. Furthermore,  $\text{var}(A) = 2\text{var}(B) = 3\text{var}(C)$ . 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.

Total Payoff  $TP = aA + bB + cC$

Variation or the total payoff can be written as

$$\text{var}(TP) = \text{var}(aA + bB + cC)$$

Expanding variance we have

$$a^2\text{var}(A) + b^2\text{var}(B) + c^2\text{var}(C) + 2ab \text{cov}(A, B) + 2bc \text{cov}(B, C) + 2ac \text{cov}(A, C)$$

We assume that the three companies are independent, therefore

$$\text{cov}(A, B) = \text{cov}(B, C) = \text{cov}(A, C) = 0$$

Since  $\text{var}(A) = 2\text{var}(B) = 3\text{var}(C)$  we can write  $\text{var}(TP)$  as:

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

Since we are given  $a + b + c = 1$  we can rewrite as:

$$\text{var}(TP) = (3a^2 + \frac{3}{2}b^2 + (1 - a - b)^2)\text{var}(C) = (4a^2 + 2ab - 2a + \frac{5}{2}b^2 - 2b + 1)\text{var}(C)$$

To minimize  $\text{var}(TP)$  we take the partial derivatives with respect to  $a$  and  $b$  and set them equal to 0:

$$\begin{aligned}\frac{\partial}{\partial a}\text{var}(TP) &= 8a + 2b - 2 \\ \frac{\partial}{\partial b}\text{var}(TP) &= 2a + 5b - 2\end{aligned}$$

Setting these two equations to 0 and solving for  $a$  and  $b$  we arrive at

$$a = \frac{1}{6}, b = \frac{1}{3}, c = \frac{1}{2}$$

## Question 3: Turtles

Next, suppose that the lifespan of a species of turtle follows a uniform distribution over  $[0, \theta]$ . Here, parameter  $\theta$  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  $y_1, y_2, \dots, y_n$ .

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

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

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

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  $\theta$  is  $\hat{\theta} = \max(x_1, x_2, \dots, x_n)$

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  $\theta$ ?

For  $n = 1$  we have  $\hat{\theta}_{ml} = y_1$ , or the value of the sample. Therefore the expectation is given by:

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

4. Is the maximum likelihood estimator biased?

Since  $E(\hat{\theta}_{ml}|\theta) = \frac{\theta}{2}$  it is a biased estimator; it does not equal  $\theta$

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

For  $n > 1$ ,  $E(\hat{\theta}_{ml}|\theta) = \int_0^\theta f(y_{max}|\theta)y_{max} dy = \int_0^\theta \frac{1}{\theta}y dy = \frac{\theta}{2}$

6. Is the maximum likelihood estimator consistent?

For very large  $n$  the MLE is always  $\frac{\theta}{2}$ , so it is not 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 endogeneity 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  $z_1$  and  $z_2$ .

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(car)
library(lmtest)
library(sandwich)
library(psych)
library(ivpack)
library(lattice)
library(dplyr)

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

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 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 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 ...

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

```

Examine the *wage* variable

```

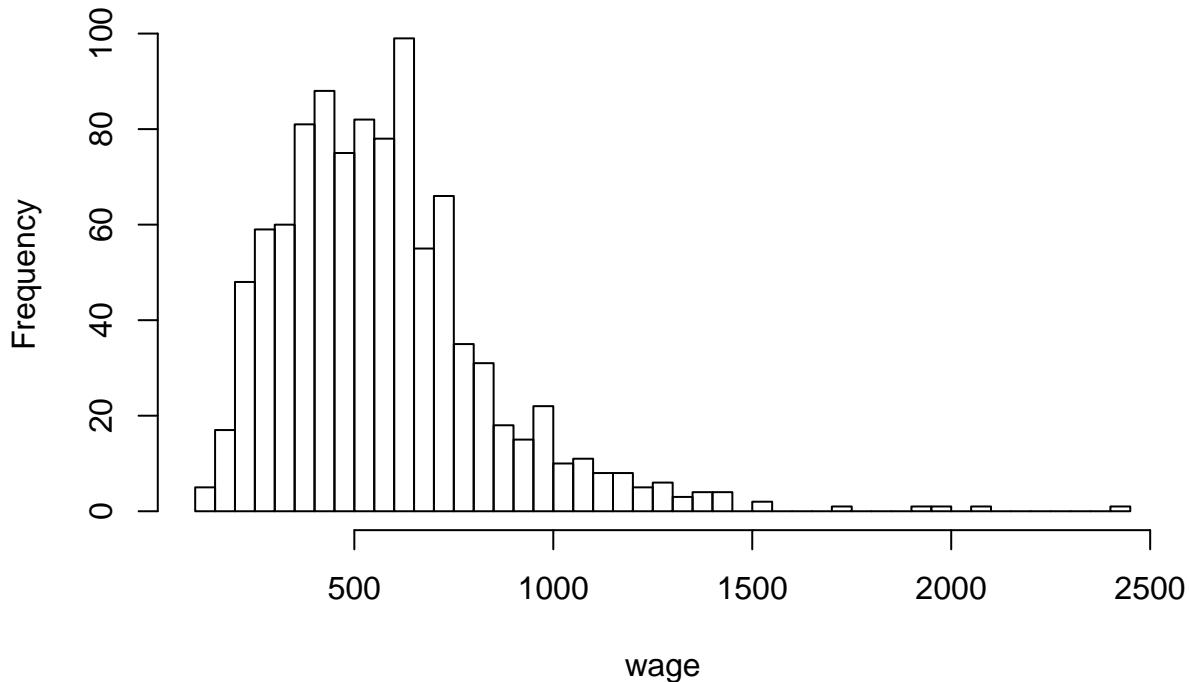
summary(df1$wage)

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

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

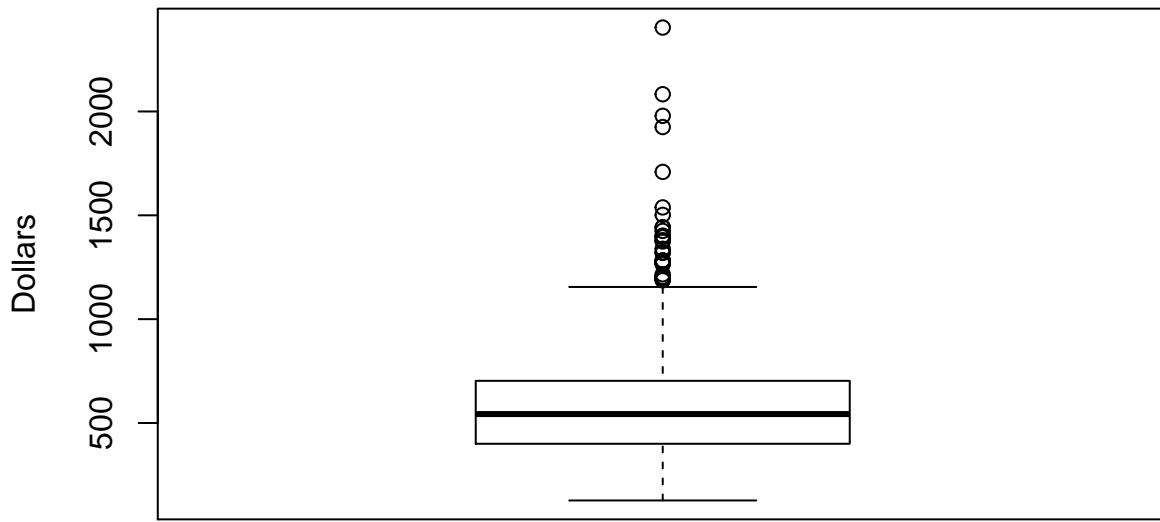
```

## Histogram of wage



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

## Box Plot of Wage



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.

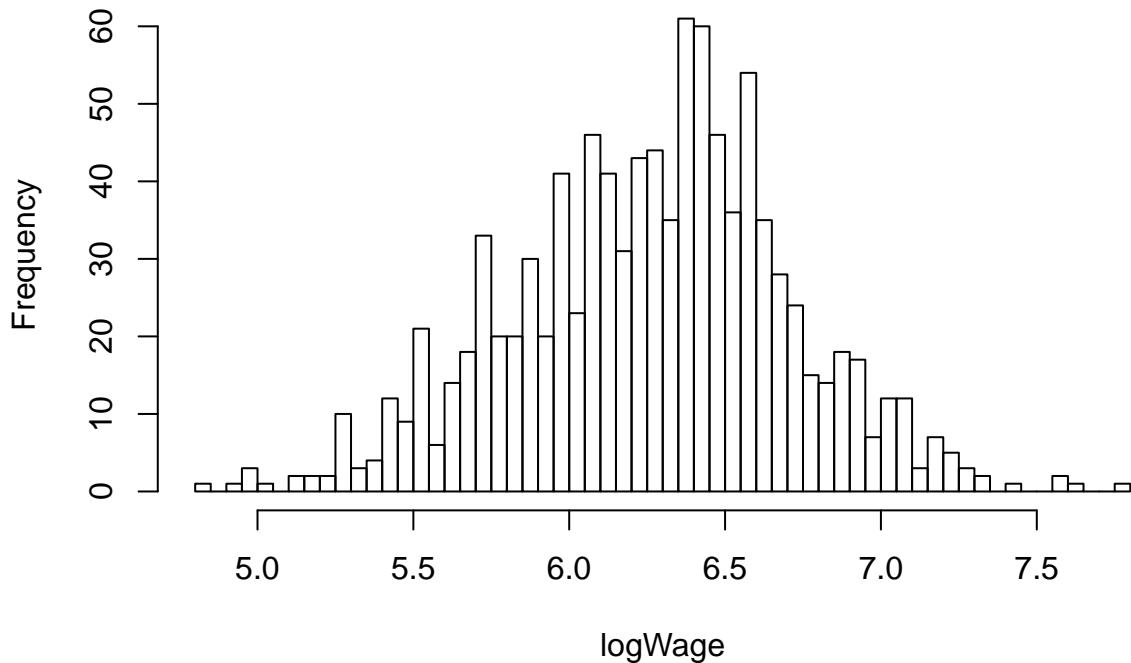
```
summary(df1$logWage)
```

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

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

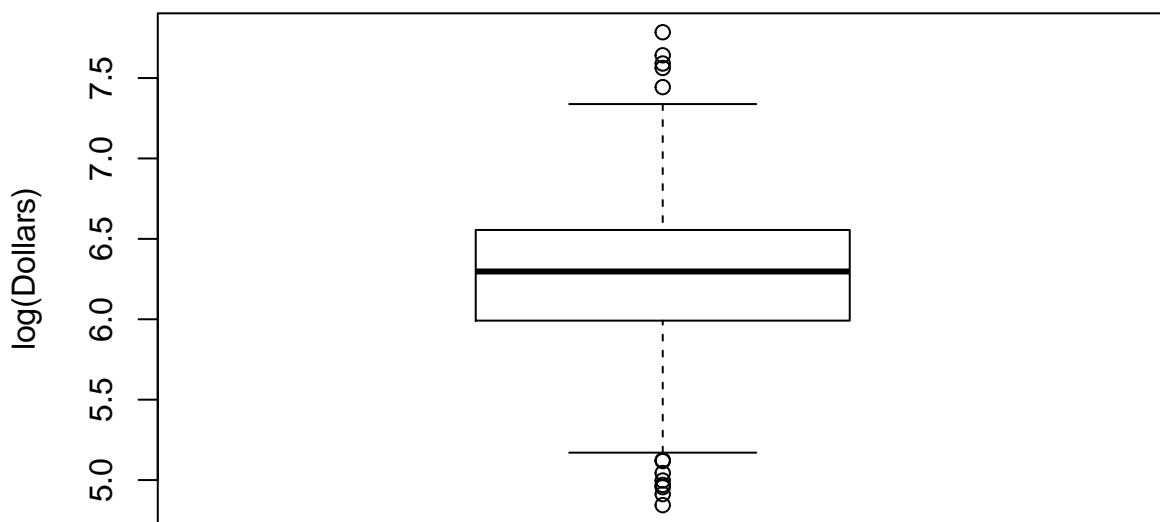
```
hist(df1$logWage, breaks=50, main='Histogram of logWage', xlab='logWage')
```

**Histogram of logWage**



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

**Box Plot of log(Wage)**



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

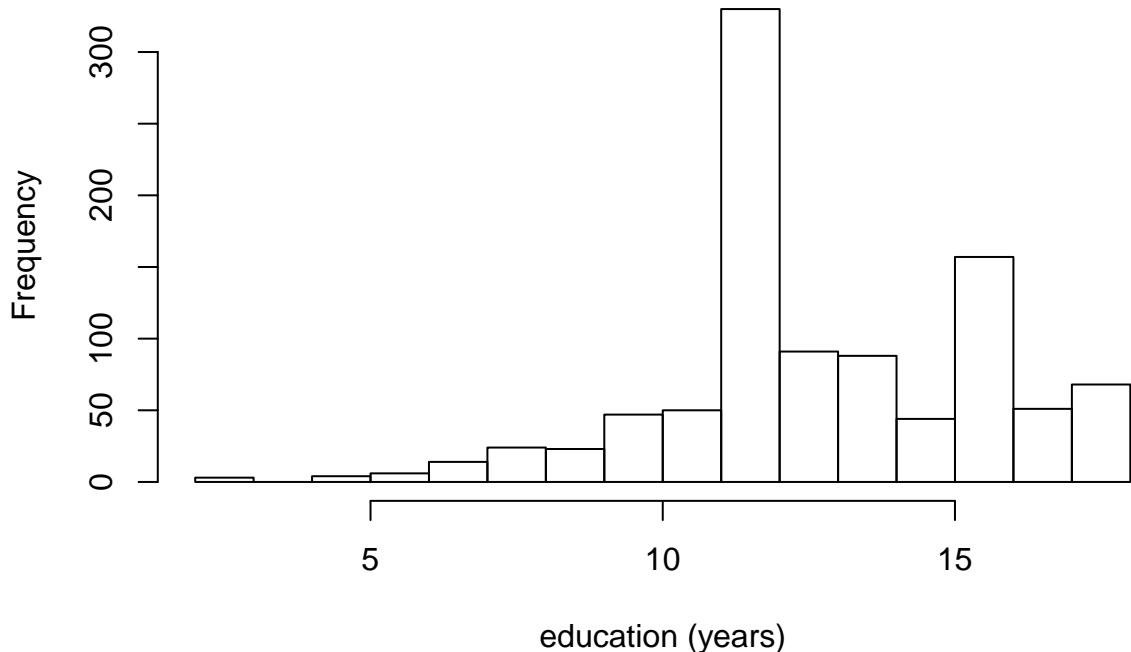
Examine the *education* variable

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

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

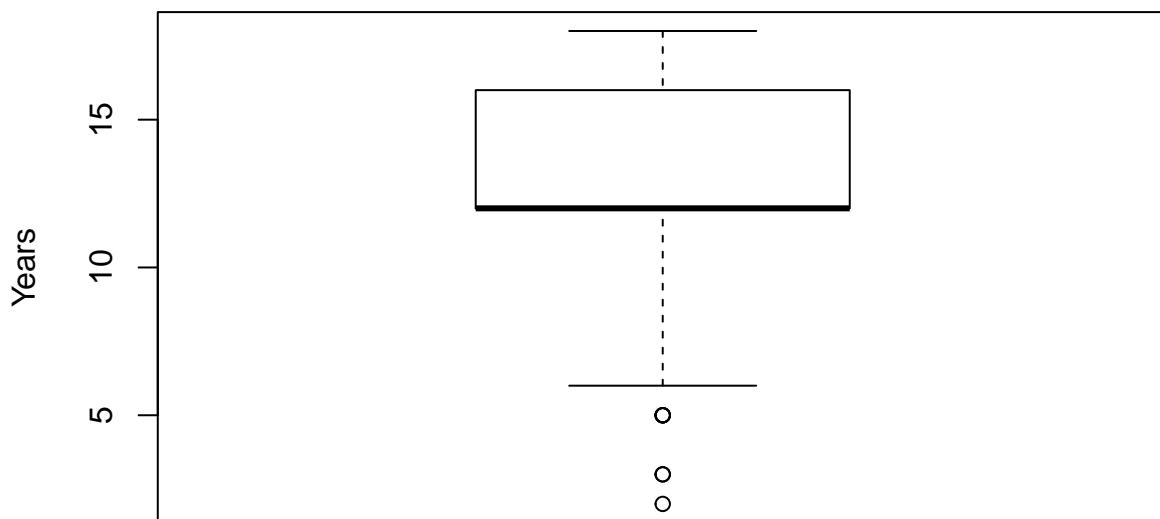
```
hist(df1$education, breaks=18, main='Histogram of education', xlab='education (years)')
```

**Histogram of education**



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

**Box Plot of Education**



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.

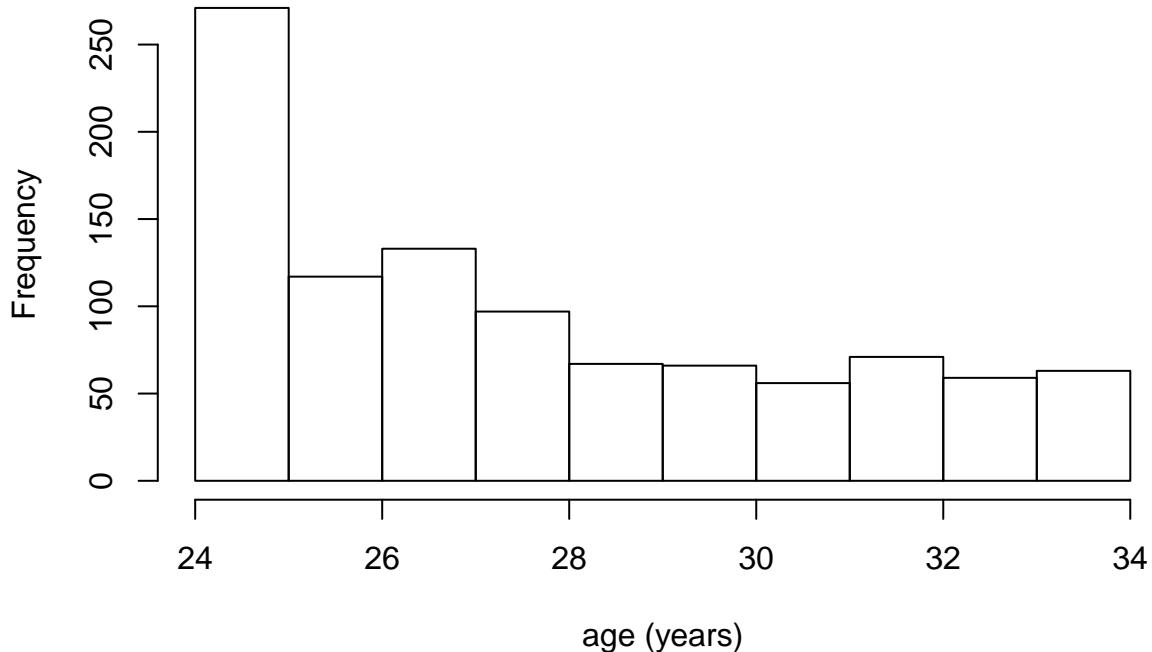
Examine the *age* variable.

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

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

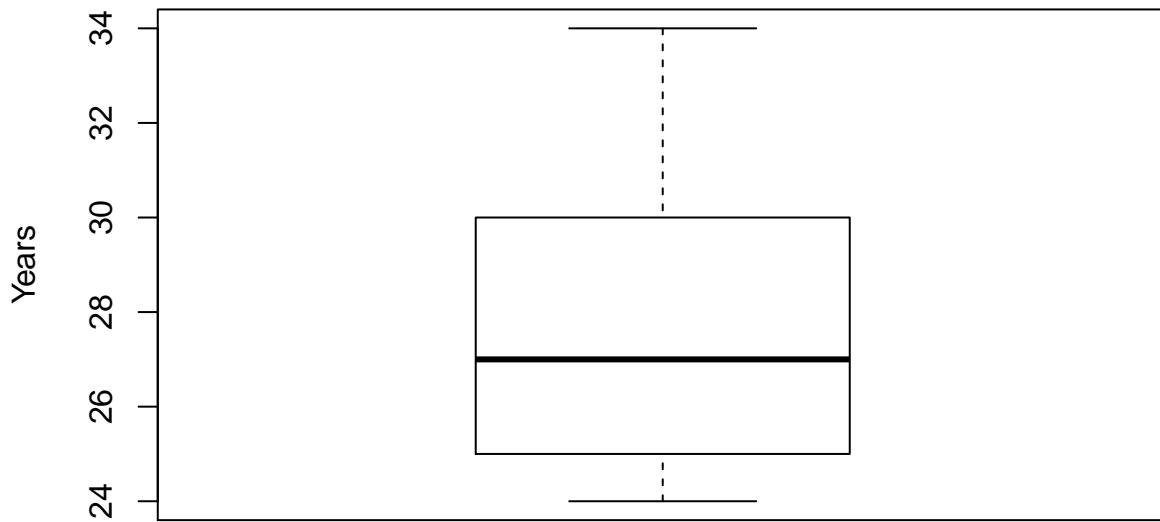
```
hist(df1$age, breaks=10, main='Histogram of age', xlab='age (years)')
```

**Histogram of age**



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

## Box Plot of Age



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-caucasian samples in the data set.

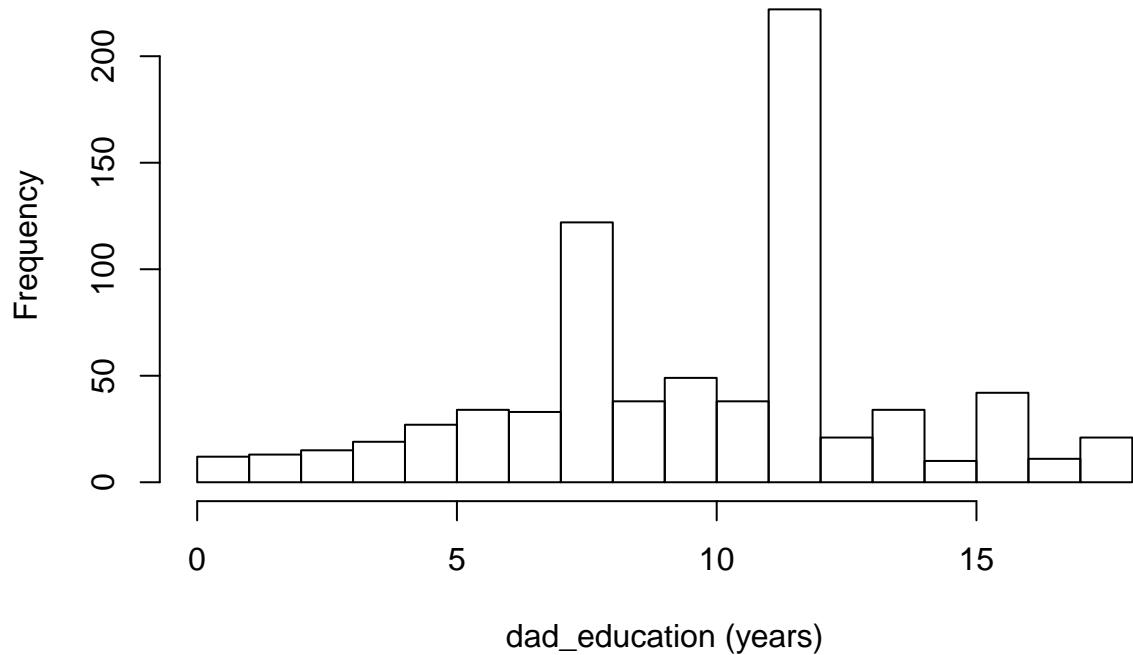
Examine the *dad\_education* variable

```
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
```

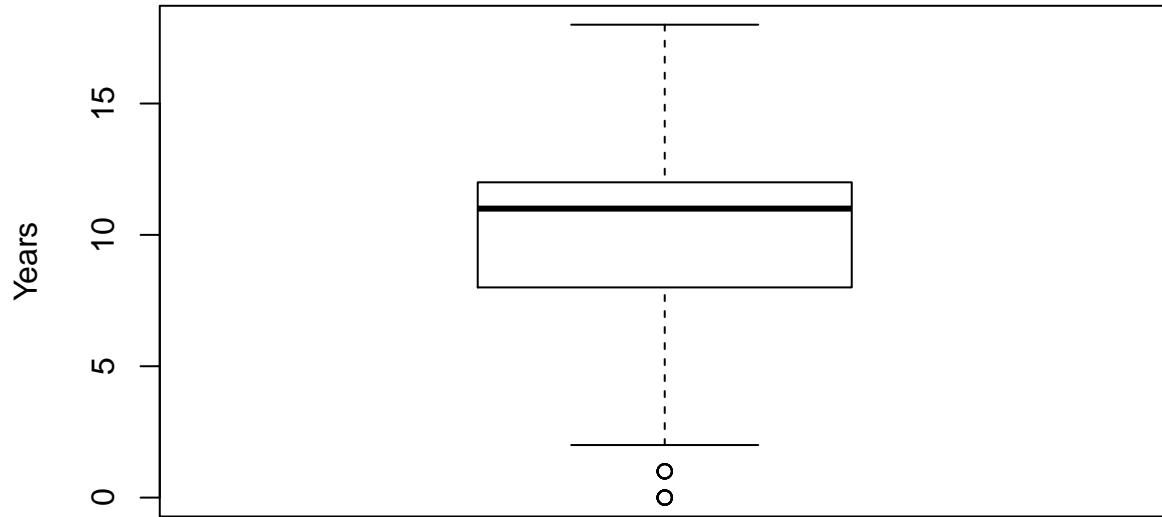
```
hist(df1$dad_education, breaks=20, main='Histogram of dad_education', xlab='dad_education (years)')
```

## Histogram of dad\_education



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

## Box Plot of Fathers Education



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 education 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.

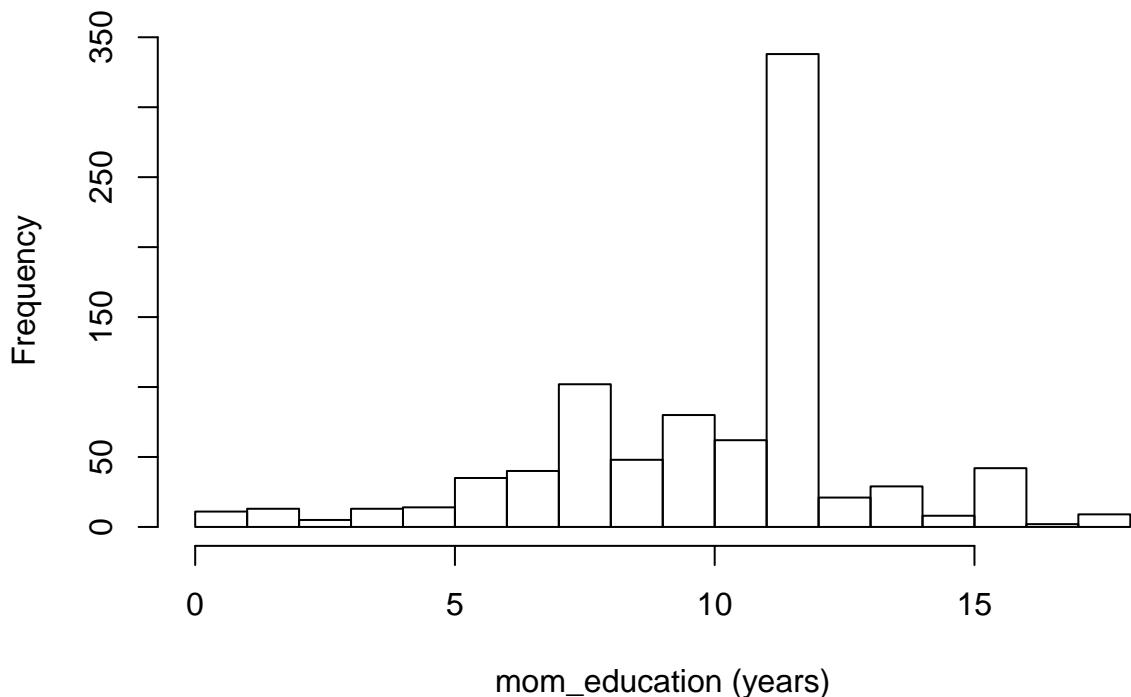
Examine *mom\_education* variable

```
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
```

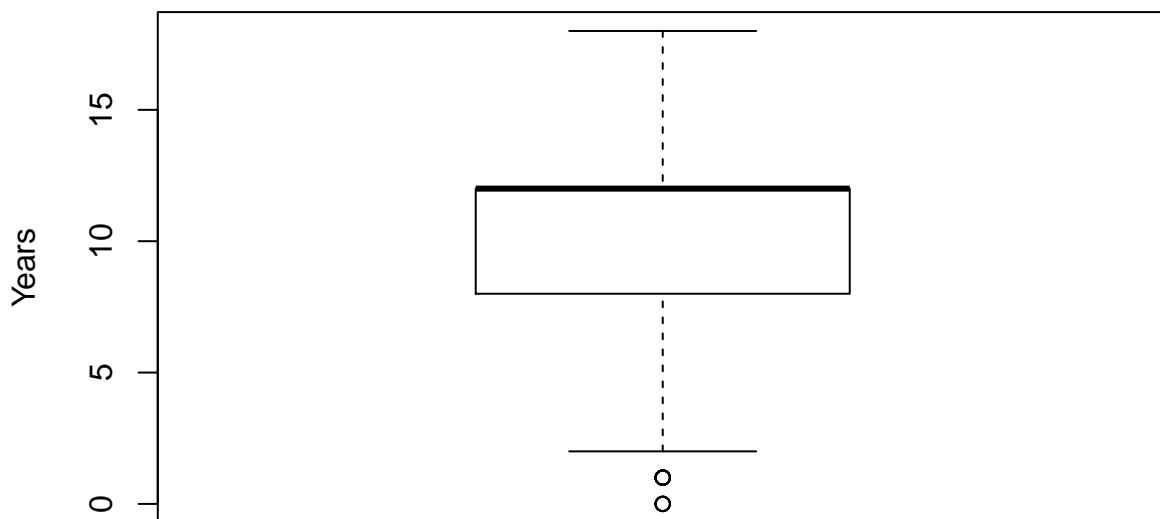
```
hist(df1$mom_education, breaks=20, main='Histogram of mom_education', xlab='mom_education (years)')
```

**Histogram of mom\_education**



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

**Box Plot of Mothers Education**



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

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

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

```
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

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

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

```
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.

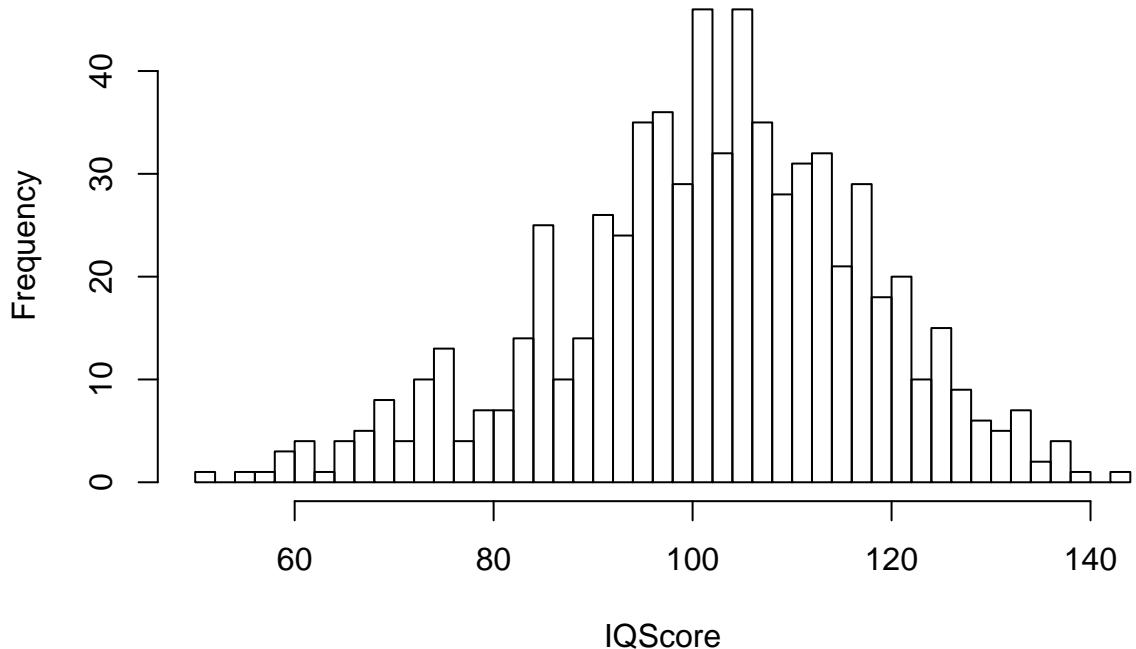
Examine *IQScore* variable

```
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
```

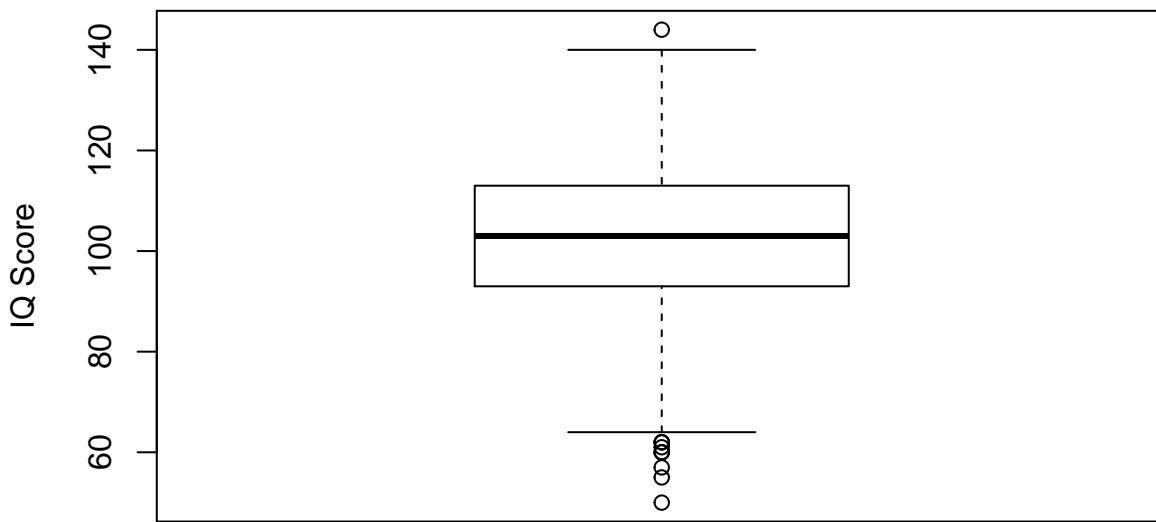
```
hist(df1$IQscore, breaks=50, main='Histogram of IQScore', xlab='IQScore')
```

## Histogram of IQScore



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

## Box Plot of 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.

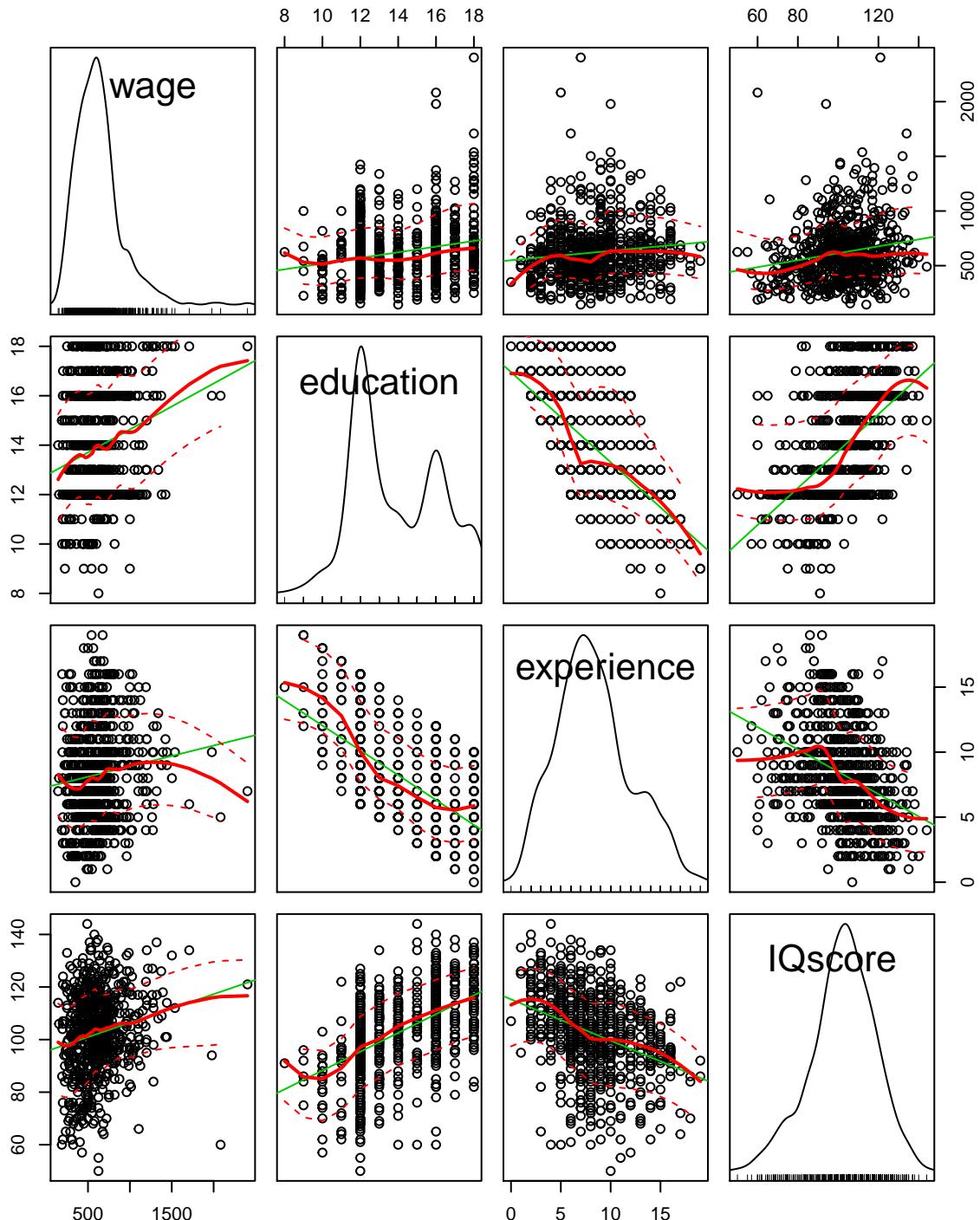
```
df1$lnWage <- log(df1$wage)
df1$experienceSquare <- df1$experience^2
```

## 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.

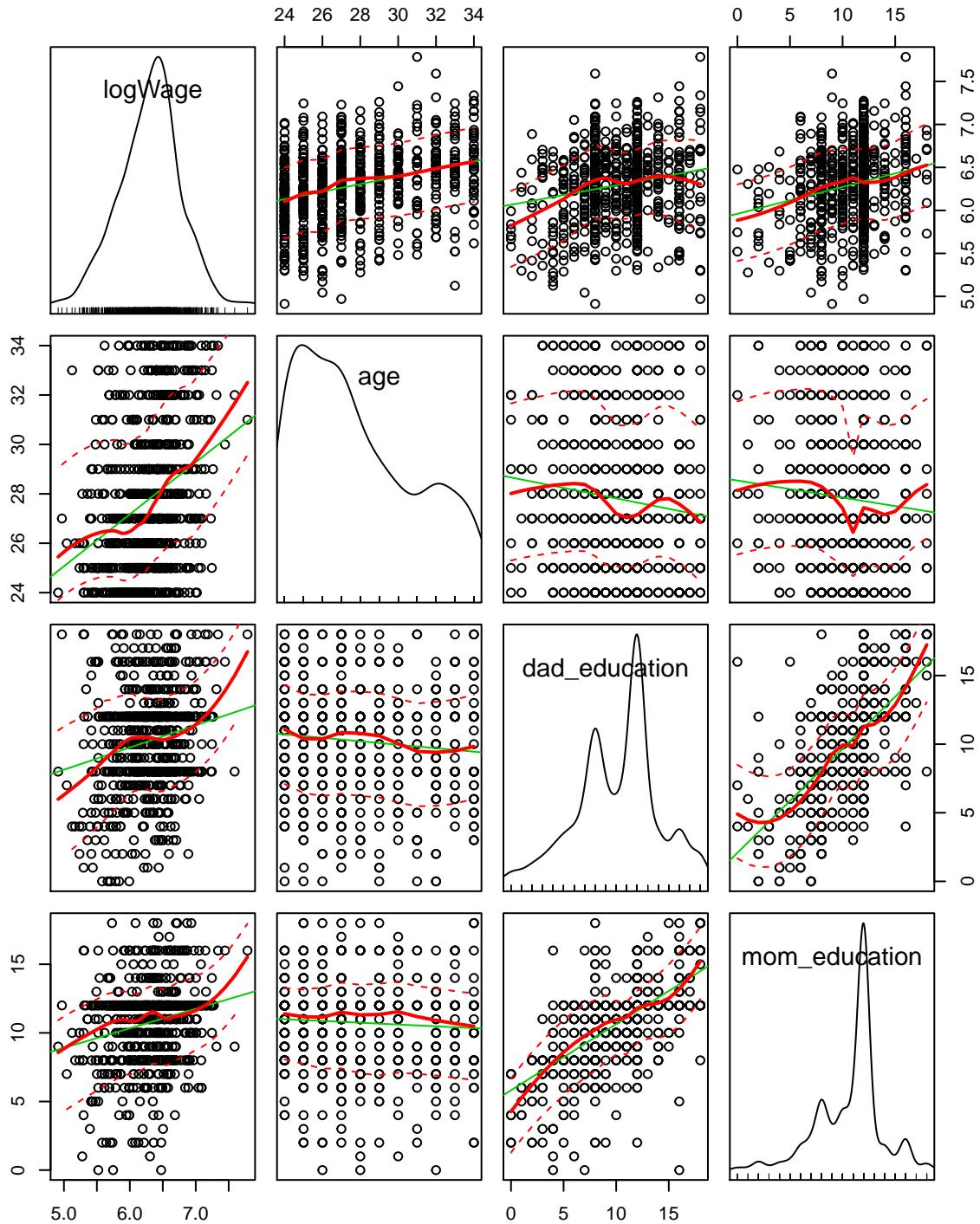
Examine *wage*, *education*, *experience*, *experienceSquare* in a scatterplot matrix

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



Examining *logWage*, *age*, *dad\_education*, *mom\_education*

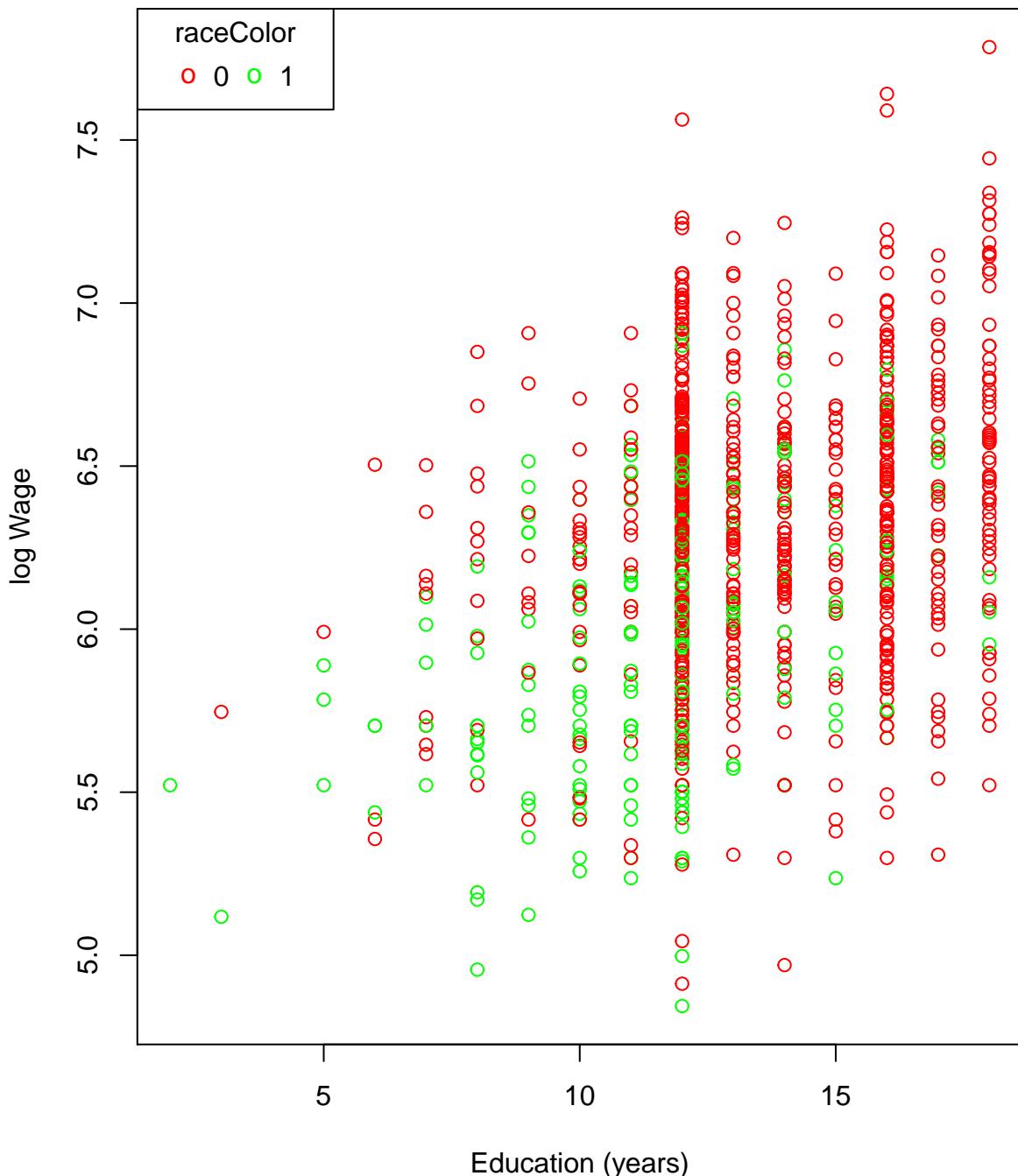
```
scatterplotMatrix(~ logWage + age + dad_education + mom_education, data=df1)
```



Exploring race effects on *education* and *wage*

```
plot(df1$education, df1$logWage, col=ifelse(df1$raceColor==0,'red','green'),
     main='Effect of Race on Education and Wage', xlab='Education (years)',
     ylab='log Wage')
legend('topleft',legend=c('0','1'),col=c('red','green'), pch='o',title='raceColor', horiz=TRUE)
```

## Effect of Race on Education and Wage



### Question 4.3

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

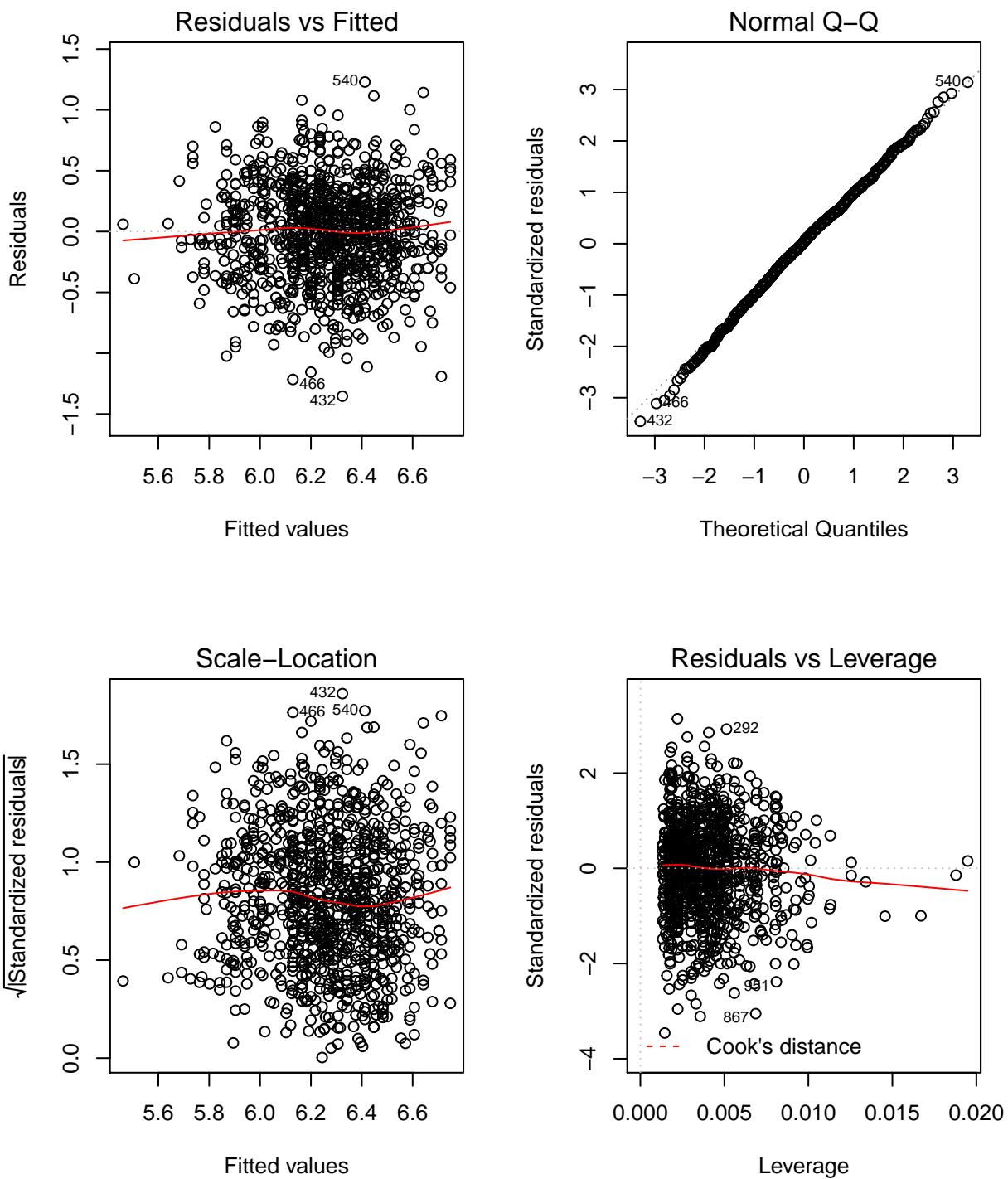
```
model4.3 <- lm(logWage ~ education + experience + age + raceColor, data=df1)
```

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
```

```
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 education, 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 education 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.

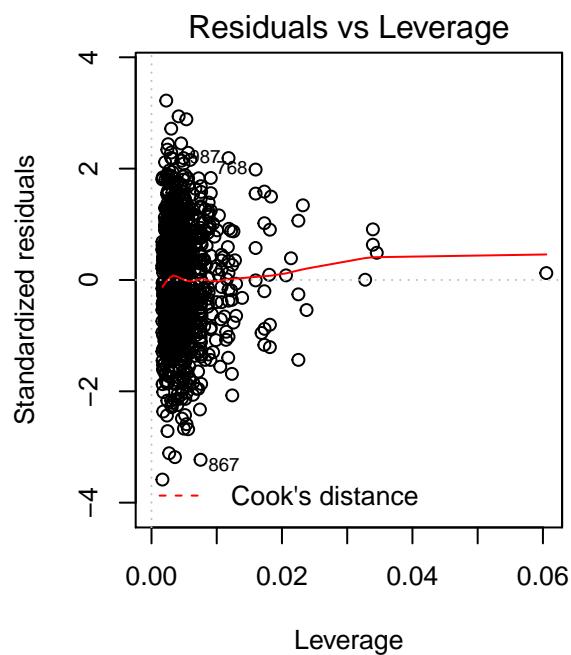
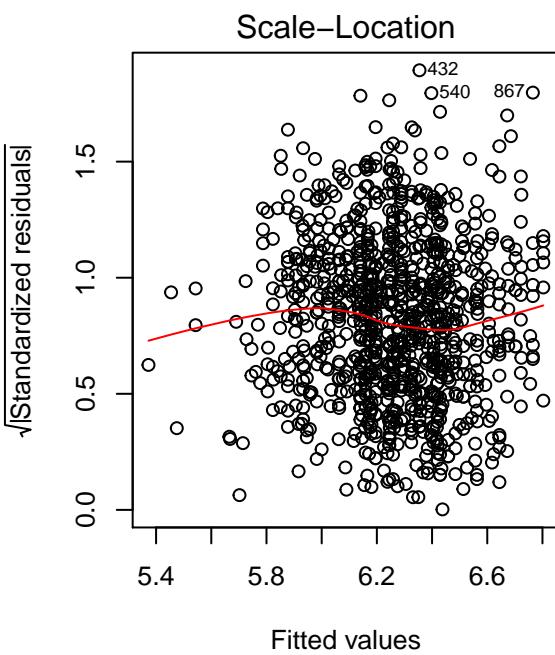
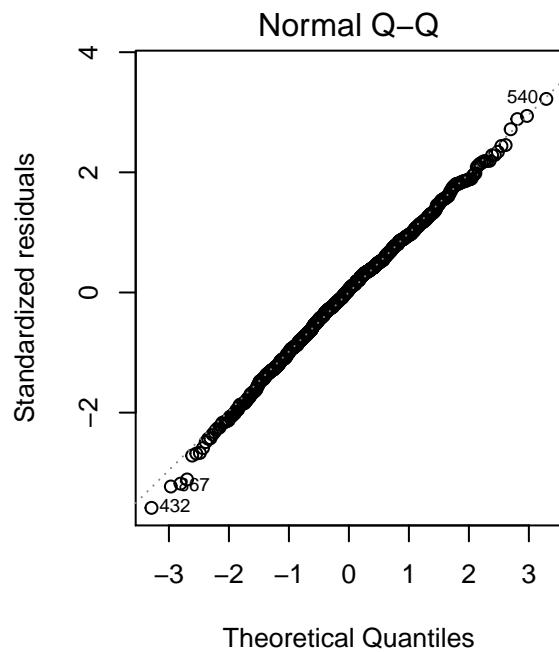
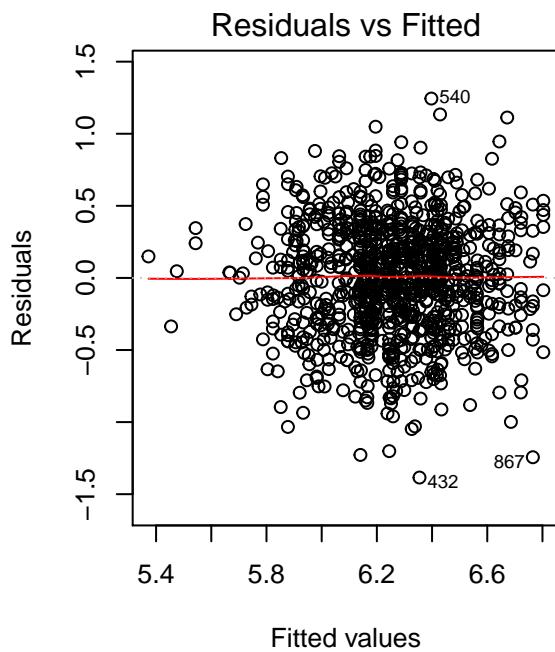
### Question 4.4

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

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

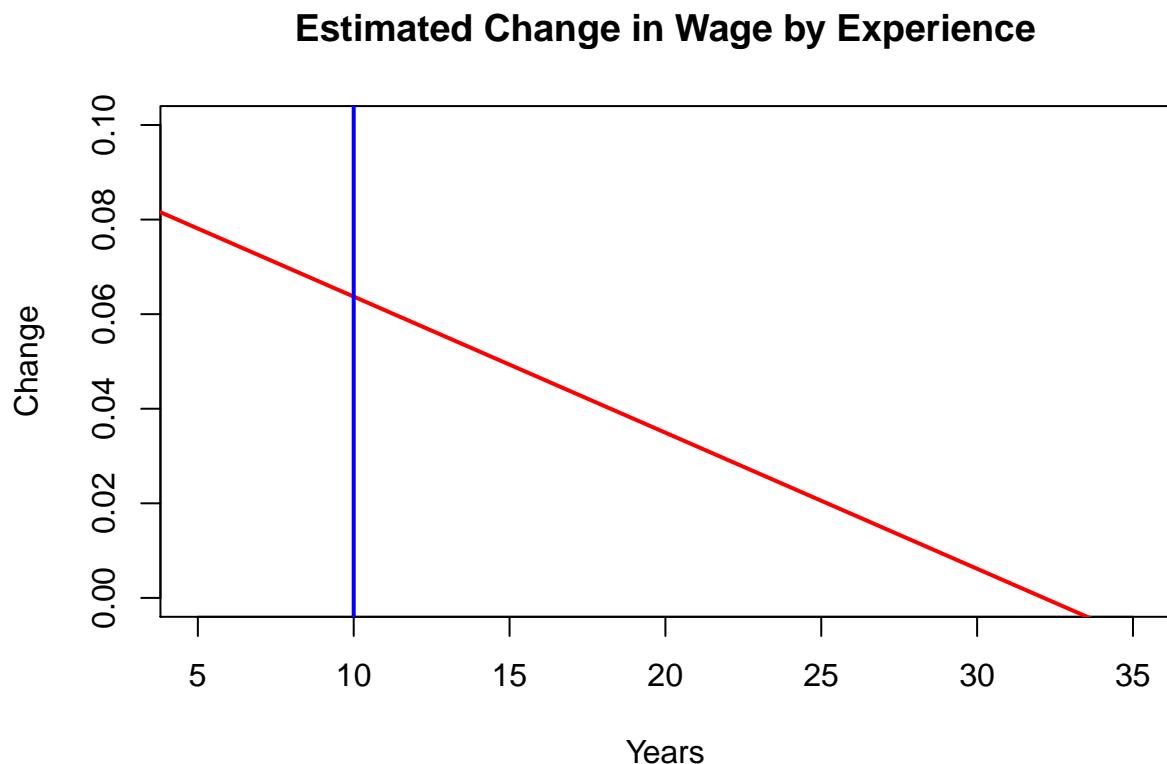
```
##
## Call:
## lm(formula = log(wage) ~ education + experience + experienceSquare +
##     raceColor, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38464 -0.25558  0.01909  0.25782  1.24410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (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
```

```
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(5,35), ylim=c(0,.1), ylab='Change',
      xlab='Years', main='Estimated Change in Wage by Experience')
abline(a = coef(model4.4)[3], b = coef(model4.4)[4], 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?

```
# Calculate % Change directly from the model parameters
100*(coef(model4.4)[3]+10*coef(model4.4))
```

```
## experience
## 6.371415
```

6.37% 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
```

```
coeftest(model4.5, vcov=vcovHC)
```

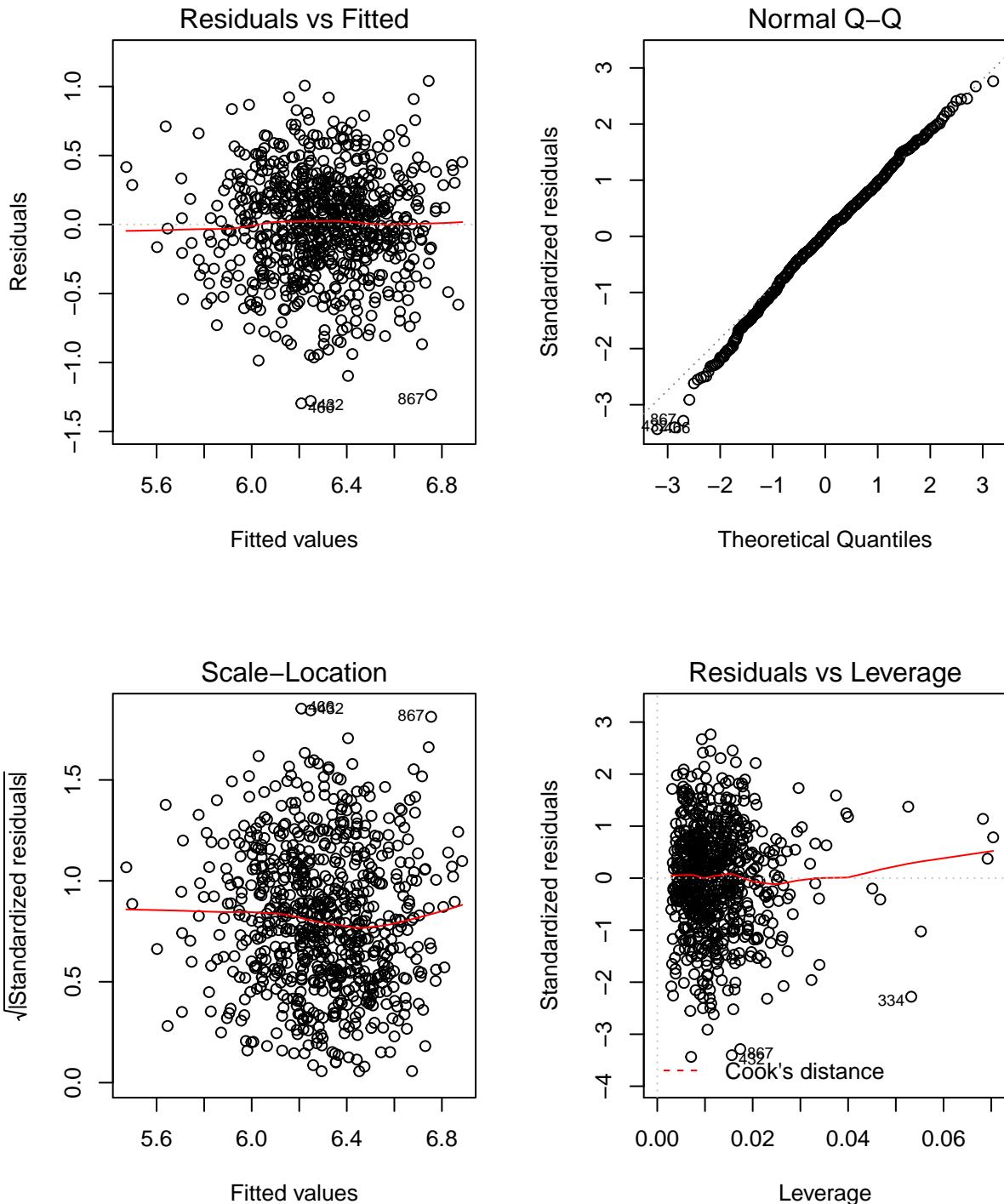
```
##
## t test of coefficients:
## 
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.64222960  0.15028402 30.8897 < 2.2e-16 ***
## education   0.06817013  0.00806946  8.4479 < 2.2e-16 ***
## experience  0.09734192  0.01320338  7.3725 4.649e-13 ***
## experienceSquare -0.00295679  0.00066974 -4.4148 1.167e-05 ***
## raceColor   -0.21302264  0.04096904 -5.1996 2.611e-07 ***
## dad_education -0.00114737  0.00568040 -0.2020  0.839984  
## mom_education 0.01131764  0.00699715  1.6175  0.106220  
## rural       -0.09193774  0.03189812 -2.8822  0.004067 ** 
## city        0.17821373  0.03191780  5.5835 3.351e-08 ***
```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```



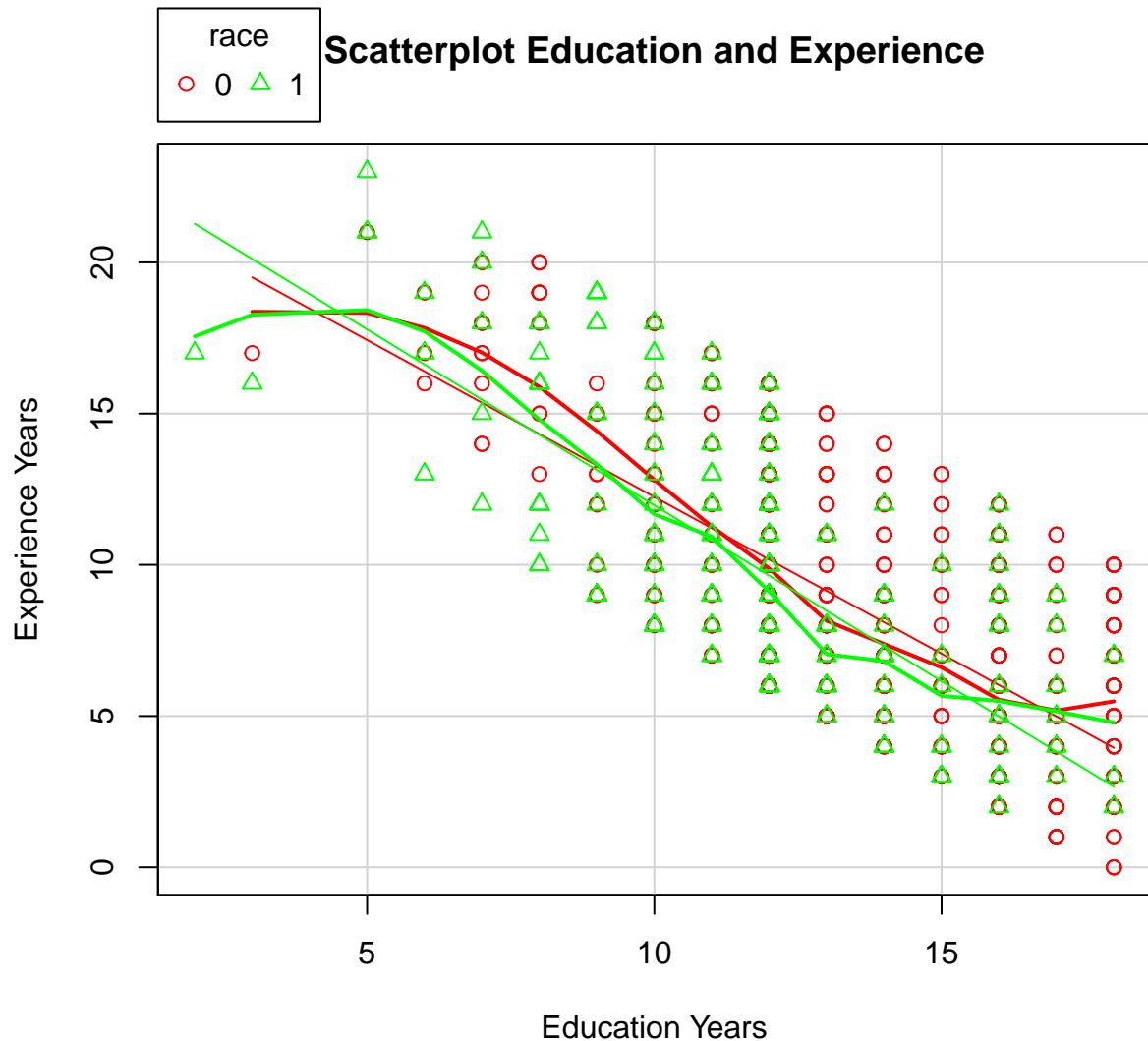
1. What are the number of observations used in this regression? Are missing values a problem? Analyze the missing values, if any, and see if there is any discernible pattern with wage, education, experience, and raceColor.

There are  $1000 - 277 = 723$  observations.

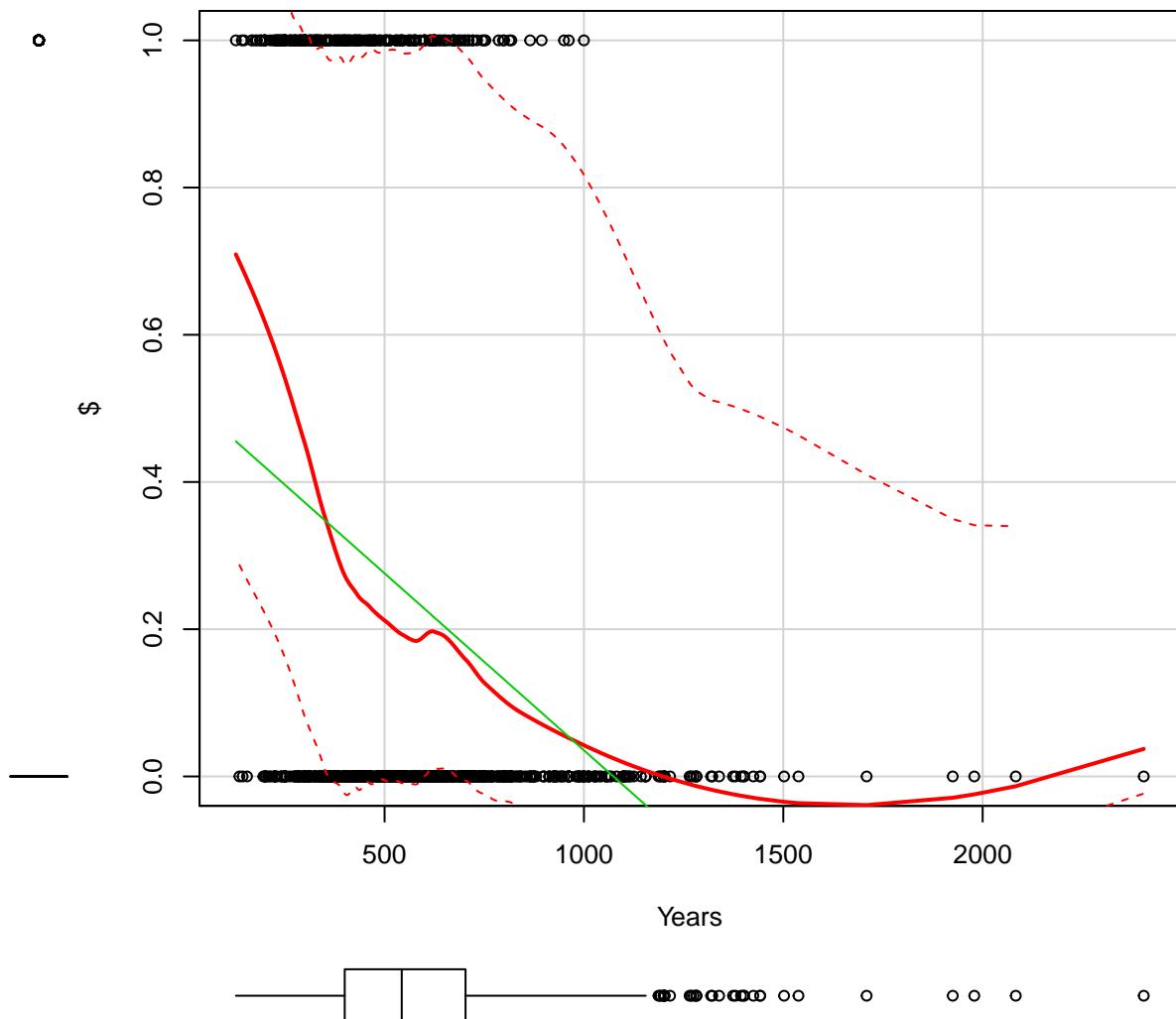
```
scatterplot(df1$education, df1$wage, groups=df1$raceColor, by.groups=TRUE,
           xlab='Years', ylab='$',
           main='Scatterplot of Wage and Education',
           legend.title='race', col=c('red','green'))
```



```
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('red','green'))
```

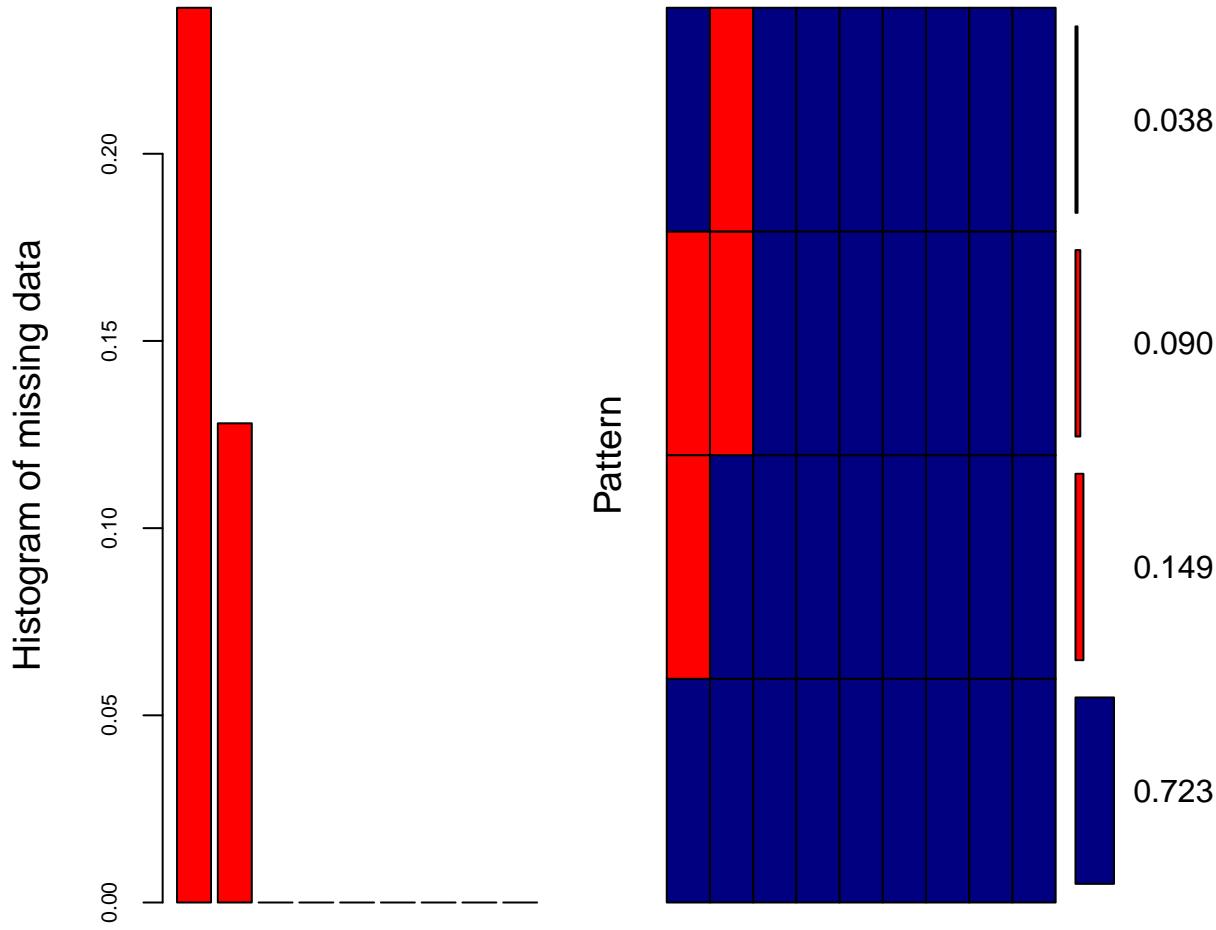


## Scatterplot Education and Experience



There is a fixed linear relationship between education and experience.

```
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 force the throwing out of the corresponding rows of the data frame, significantly reducing the power of the model.

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.

```
# replace the NA values in the dad_education and mom_education with respective mean values
df1.na <- df1
df1.na$dad_education <- ifelse(is.na(df1.na$dad_education),
                                 mean(df1$dad_education, na.rm=TRUE),
                                 df1.na$dad_education)
df1.na$mom_education <- ifelse(is.na(df1.na$mom_education),
                                 mean(df1$mom_education, na.rm=TRUE),
                                 df1.na$mom_education)
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

summary(df1.na$dad_education)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      0.00   8.00  10.18    10.18  12.00   18.00

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

summary(df1.na$mom_education)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      0.00   9.00  11.00    10.45  12.00   18.00

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

##
## Call:
## lm(formula = logWage ~ education + experience + experienceSquare +
##     raceColor + dad_education + mom_education + rural + city,
##     data = df1.na)
##
## Residuals:
##      Min       1Q       Median       3Q       Max
## -1.30741 -0.23286  0.01943  0.24786  1.28807
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```

## (Intercept)      4.729e+00  1.226e-01  38.584 < 2e-16 ***
## education       7.097e-02  6.499e-03  10.920 < 2e-16 ***
## experience      8.958e-02  1.124e-02   7.970 4.36e-15 ***
## experienceSquare -2.678e-03  5.318e-04  -5.036 5.65e-07 ***
## raceColor        -2.313e-01  3.099e-02  -7.464 1.84e-13 ***
## dad_education    -3.513e-05  4.416e-03  -0.008 0.993656
## mom_education    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

```

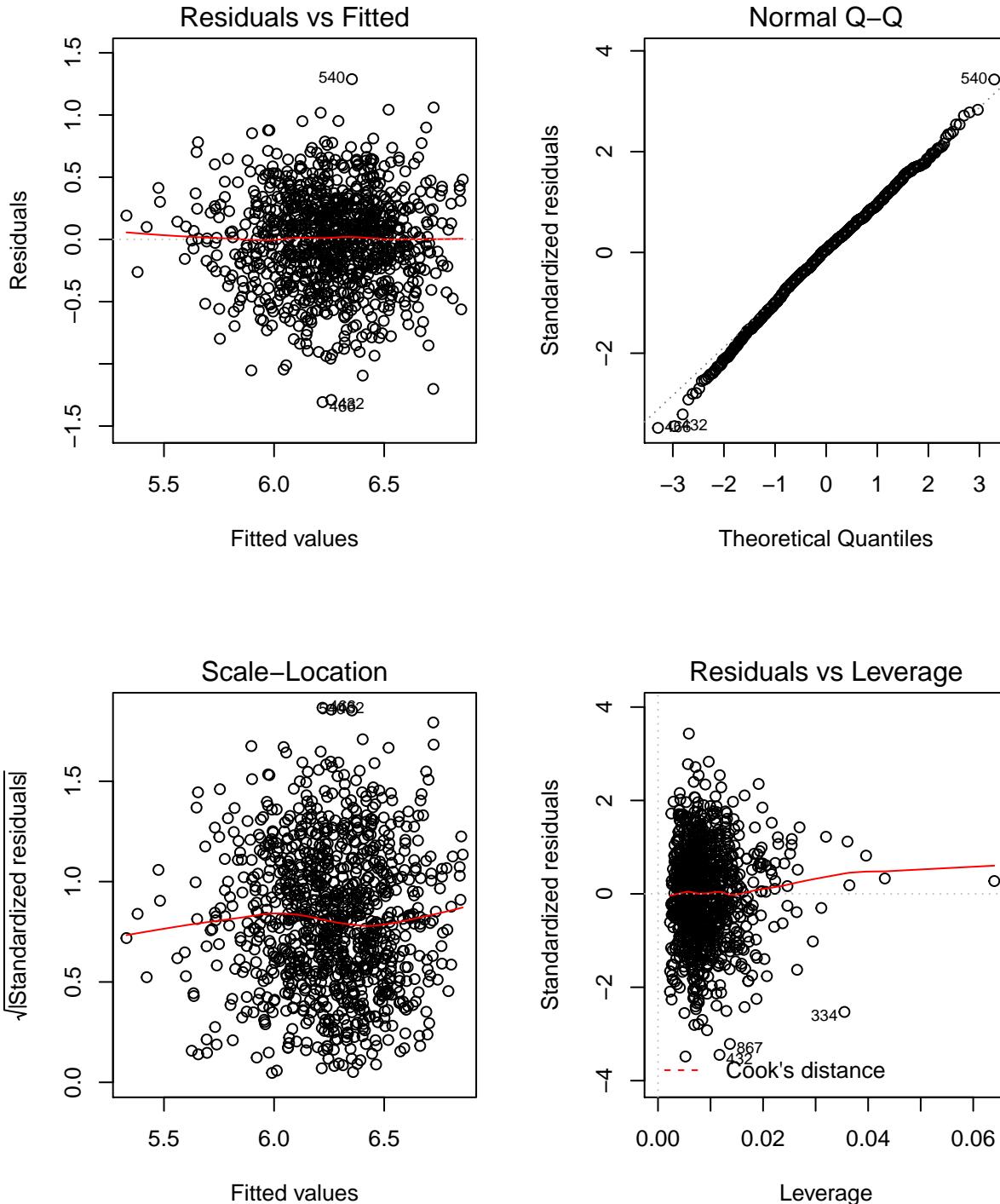
```
coefest(model4.5.na, vcov=vcovHC)
```

```

##
## t test of coefficients:
##
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.7292e+00  1.2755e-01 37.0779 < 2.2e-16 ***
## education       7.0966e-02  6.6074e-03 10.7403 < 2.2e-16 ***
## experience      8.9585e-02  1.0938e-02  8.1900 7.998e-16 ***
## experienceSquare -2.6782e-03  5.1840e-04 -5.1663 2.886e-07 ***
## raceColor        -2.3132e-01  3.0655e-02 -7.5460 1.015e-13 ***
## dad_education    -3.5127e-05  4.8214e-03 -0.0073 0.9941884
## mom_education    3.4848e-03  5.3451e-03  0.6520 0.5145786
## rural            -9.5287e-02  2.6912e-02 -3.5407 0.0004176 ***
## city              1.6714e-01  2.6395e-02   6.3322 3.658e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
par(mfrow = c(2,2))
plot(model4.5.na, sub.caption="Model Diagnostic Plots - NA replaced with mean")
```



Regressing against a modified *dad\_education* and *mom\_education* that has the mean of the variable in place of NA does not improve the model. The estimates of the two coefficients remains highly statistically insignificant.

- 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?

```

# make a temporary datafrome that doesn't have IQscore
df1.na <- df1[, !(names(df1.na) %in% c('IQscore'))]
# use the mice package to impute the missing data
df1.imputed <- mice(df1.na,m=5, method='pmm',
                     maxit=50, seed=500, printFlag = FALSE)
# check the results
summary(df1.imputed)

## Multiply imputed data set
## Call:
## mice(data = df1.na, m = 5, method = "pmm", maxit = 50, printFlag = FALSE,
##       seed = 500)
## Number of multiple imputations: 5
## Missing cells per column:
##          X        wage      education      experience
##          0         0           0             0
##          age      raceColor   dad_education mom_education
##          0         0           239            128
##          rural     city        z1              z2
##          0         0           0               0
##          logWage    lnWage  experienceSquare
##          0         0           0               0
## Imputation methods:
##          X        wage      education      experience
##          "pmm"    "pmm"      "pmm"        "pmm"
##          age      raceColor   dad_education mom_education
##          "pmm"    "pmm"      "pmm"        "pmm"
##          rural     city        z1              z2
##          "pmm"    "pmm"      "pmm"        "pmm"
##          logWage    lnWage  experienceSquare
##          "pmm"    "pmm"      "pmm"
## VisitSequence:
## dad_education mom_education
##          7           8
## PredictorMatrix:
##          X wage education experience age raceColor dad_education
##          X  0   0       0        0   0       0       0
##          wage 0   0       0        0   0       0       0
##          education 0   0       0        0   0       0       0
##          experience 0   0       0        0   0       0       0
##          age 0   0       0        0   0       0       0
##          raceColor 0   0       0        0   0       0       0
##          dad_education 1   1       1        1   1       1       0
##          mom_education 1   1       1        1   1       1       1
##          rural 0   0       0        0   0       0       0
##          city 0   0       0        0   0       0       0
##          z1 0   0       0        0   0       0       0
##          z2 0   0       0        0   0       0       0
##          logWage 0   0       0        0   0       0       0
##          lnWage 0   0       0        0   0       0       0
##          experienceSquare 0   0       0        0   0       0       0
##          mom_education rural city z1 z2 logWage lnWage
##          X                 0   0   0   0   0   0   0

```

```

## wage 0 0 0 0 0 0 0
## education 0 0 0 0 0 0 0
## experience 0 0 0 0 0 0 0
## age 0 0 0 0 0 0 0
## raceColor 0 0 0 0 0 0 0
## dad_education 1 1 1 1 1 1 0
## mom_education 0 1 1 1 1 1 0
## rural 0 0 0 0 0 0 0
## city 0 0 0 0 0 0 0
## z1 0 0 0 0 0 0 0
## z2 0 0 0 0 0 0 0
## logWage 0 0 0 0 0 0 0
## lnWage 0 0 0 0 0 0 0
## experienceSquare 0 0 0 0 0 0 0
## experienceSquare
## X 0
## wage 0
## education 0
## experience 0
## age 0
## raceColor 0
## dad_education 1
## mom_education 1
## rural 0
## city 0
## z1 0
## z2 0
## logWage 0
## lnWage 0
## experienceSquare 0
## Random generator seed value: 500

```

```
summary(df1.imputed)
```

```

## Multiply imputed data set
## Call:
## mice(data = df1.na, m = 5, method = "pmm", maxit = 50, printFlag = FALSE,
##       seed = 500)
## Number of multiple imputations: 5
## Missing cells per column:
##          X      wage      education      experience
##          0        0        0        0
##          age    raceColor    dad_education    mom_education
##          0        0        239        128
##          rural     city        z1        z2
##          0        0        0        0
##          logWage    lnWage  experienceSquare
##          0        0        0
## Imputation methods:
##          X      wage      education      experience
## "pmm"    "pmm"    "pmm"    "pmm"
##          age    raceColor    dad_education    mom_education
## "pmm"    "pmm"    "pmm"    "pmm"
##          rural     city        z1        z2

```

```

##          "pmm"          "pmm"          "pmm"          "pmm"
##      logWage      lnWage experienceSquare
##          "pmm"          "pmm"          "pmm"
## VisitSequence:
## dad_education mom_education
##           7           8
## PredictorMatrix:
##          X wage education experience age raceColor dad_education
## X       0   0        0        0   0       0       0
## wage    0   0        0        0   0       0       0
## education 0   0        0        0   0       0       0
## experience 0   0        0        0   0       0       0
## age     0   0        0        0   0       0       0
## raceColor 0   0        0        0   0       0       0
## dad_education 1   1        1        1   1       1       0
## mom_education 1   1        1        1   1       1       1
## rural    0   0        0        0   0       0       0
## city     0   0        0        0   0       0       0
## z1      0   0        0        0   0       0       0
## z2      0   0        0        0   0       0       0
## logWage  0   0        0        0   0       0       0
## lnWage   0   0        0        0   0       0       0
## experienceSquare 0   0        0        0   0       0       0
##          mom_education rural city z1 z2 logWage lnWage
## X         0   0        0   0   0       0       0
## wage     0   0        0   0   0       0       0
## education 0   0        0   0   0       0       0
## experience 0   0        0   0   0       0       0
## age     0   0        0   0   0       0       0
## raceColor 0   0        0   0   0       0       0
## dad_education 1   1        1   1   1       1       0
## mom_education 0   1        1   1   1       1       0
## rural    0   0        0   0   0       0       0
## city     0   0        0   0   0       0       0
## z1      0   0        0   0   0       0       0
## z2      0   0        0   0   0       0       0
## logWage  0   0        0   0   0       0       0
## lnWage   0   0        0   0   0       0       0
## experienceSquare 0   0        0   0   0       0       0
##          experienceSquare
## X         0
## wage     0
## education 0
## experience 0
## age     0
## raceColor 0
## dad_education 1
## mom_education 1
## rural    0
## city     0
## z1      0
## z2      0
## logWage  0
## lnWage   0

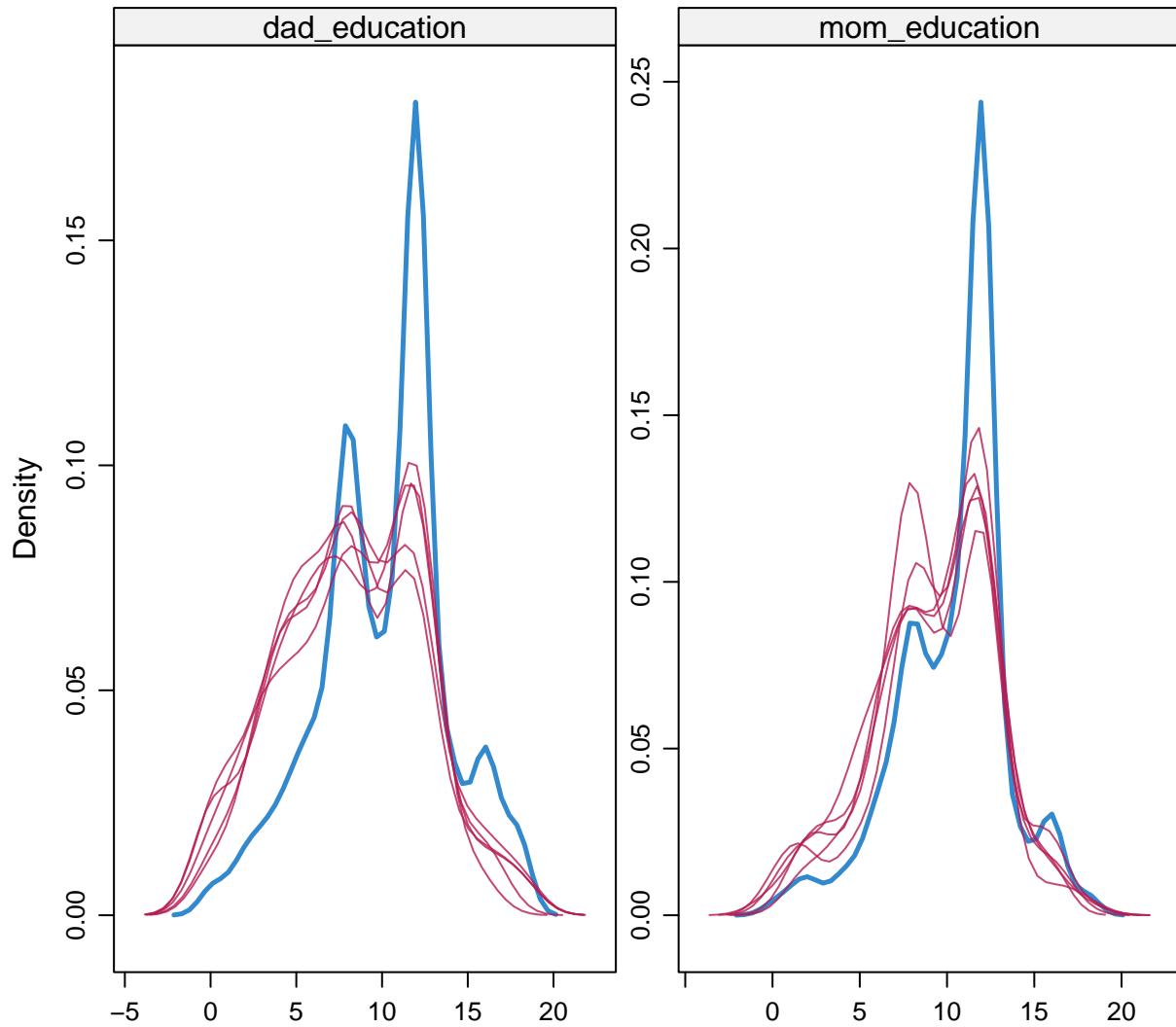
```

```

## experienceSquare          0
## Random generator seed value:  500

# density plot should show the imputed distribution to be similar to the observed
densityplot(df1.imputed)

```



```

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

```

```

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

```

```
coefest(model4.5.na, vcov=vcovHC)
```

```

## 
## t test of coefficients:
## 
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           4.64222960 0.15028402 30.8897 < 2.2e-16 ***
## education            0.06817013 0.00806946  8.4479 < 2.2e-16 ***  
## experience          0.09734192 0.01320338  7.3725 4.649e-13 ***  
## experienceSquare   -0.00295679 0.00066974 -4.4148 1.167e-05 ***  
## raceColor            -0.21302264 0.04096904 -5.1996 2.611e-07 ***  
## dad_education       -0.00114737 0.00568040 -0.2020  0.839984  
## mom_education        0.01131764 0.00699715  1.6175  0.106220  
## rural                -0.09193774 0.03189812 -2.8822  0.004067 **  
## city                 0.17821373 0.03191780  5.5835 3.351e-08 ***  
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Again the estimated coefficients remain statistically insignificant.

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

```
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
```

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

```
##  
## Call:  
## lm(formula = log(wage) ~ education + experience + experienceSquare +  
##      raceColor, data = df1)  
##  
## 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
```

```
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  $\text{cov}(z_1, u) = 0$  and that if the variable for which we want to use  $z_i$  as an indicator is  $x$  then  $\text{cov}(x, z_1) \neq 0$

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  $z_1$  and  $z_2$  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

Note: no descriptive information was provided with this data set but the data set is found to contain the following variables:

Variable	Description
<i>region</i>	region of the country
<i>urb</i>	percentage of population in urban areas
<i>lit</i>	literacy rate of district
<i>voteshare</i>	unknown exactly what this is - margin of victory?
<i>absolute_wealth</i>	candidate wealth

```
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)

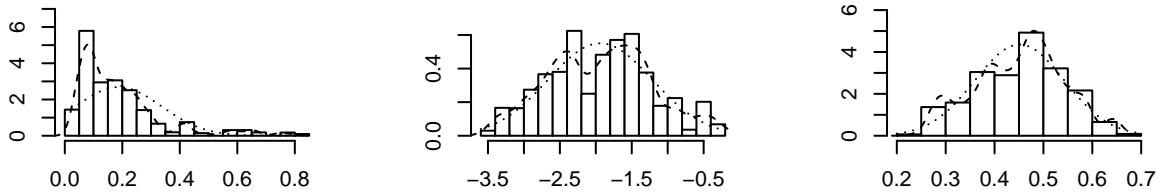
##           X              region            urb             lit
## Min.   : 1.0   Region 1:1183   Min.   :0.02835   Min.   :0.2418
## 1st Qu.: 625.2  Region 2: 690   1st Qu.:0.08387   1st Qu.:0.3846
## Median :1249.5  Region 3: 625   Median :0.14657   Median :0.4602
## Mean   :1249.5                    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
```

```

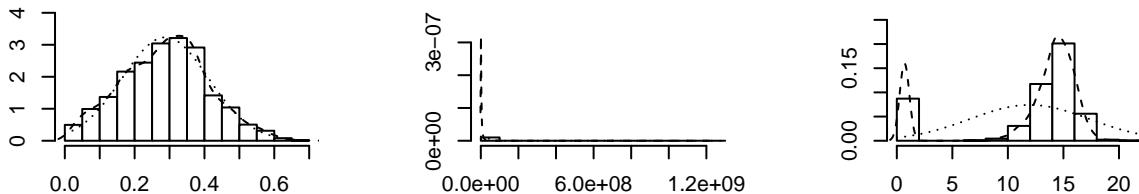
df2$logWealth <- log(df2$absolute_wealth)
df2$logUrb<- log(df2$urb)
par(mar=c(3,3,3,3))
mhist <- df2[,c('urb','logUrb','lit','voteshare','absolute_wealth', 'logWealth')]
mhist$region <- as.numeric(df2$region)
multi.hist(mhist)

```

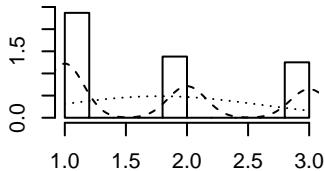
### histogram, Density, and Normal Fhistogram, Density, and Normal F



### histogram, Density, and Normal Fhistogram, Density, and Normal F



### histogram, Density, and Normal F



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 examining 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.

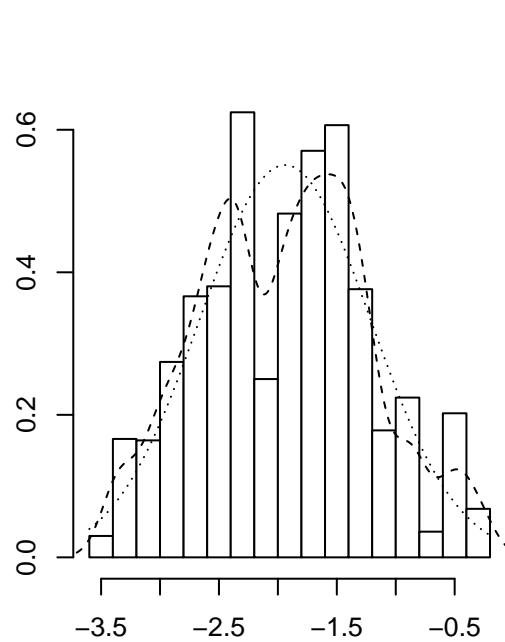
After a bit of research we found that the 2.0 value is not a mistake but the actual data. It is the minimum wealth recorded for those candidates so we leave the data as it is.

```
# rename the absolute_wealth variable for convenience
df2$wealth<-df2$absolute_wealth
df2$logWealth <- log(df2$wealth)
summary(df2)
```

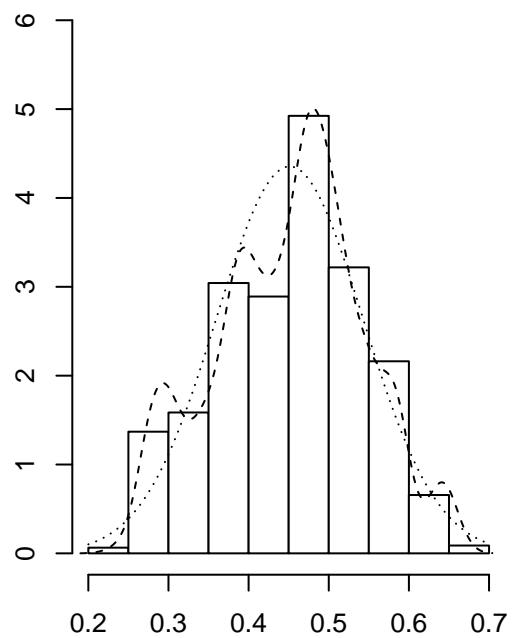
```
##           X             region            urb            lit
##  Min.   : 1.0   Region 1:1183   Min.   :0.02835   Min.   :0.2418
##  1st Qu.: 625.2  Region 2: 690   1st Qu.:0.08387   1st Qu.:0.3846
##  Median :1249.5  Region 3: 625   Median :0.14657   Median :0.4602
##  Mean   :1249.5                   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
##  Min.   :0.006037   Min.   :2.000e+00   Min.   : 0.6931
##  1st Qu.:0.199620   1st Qu.:1.875e+05   1st Qu.:12.1415
##  Median :0.293398   Median :1.337e+06   Median :14.1057
##  Mean   :0.287860   Mean   :5.034e+06   Mean   :11.9606
##  3rd Qu.:0.367978   3rd Qu.:4.092e+06   3rd Qu.:15.2245
##  Max.   :0.693324   Max.   :1.216e+09   Max.   :20.9192
##           NA's   :1                 NA's   :1
##
##           logUrb          wealth
##  Min.   :-3.5632   Min.   :2.000e+00
##  1st Qu.:-2.4785   1st Qu.:1.875e+05
##  Median :-1.9202   Median :1.337e+06
##  Mean   :-1.9387   Mean   :5.034e+06
##  3rd Qu.:-1.4139   3rd Qu.:4.092e+06
##  Max.   :-0.2202   Max.   :1.216e+09
##           NA's   :1
```

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

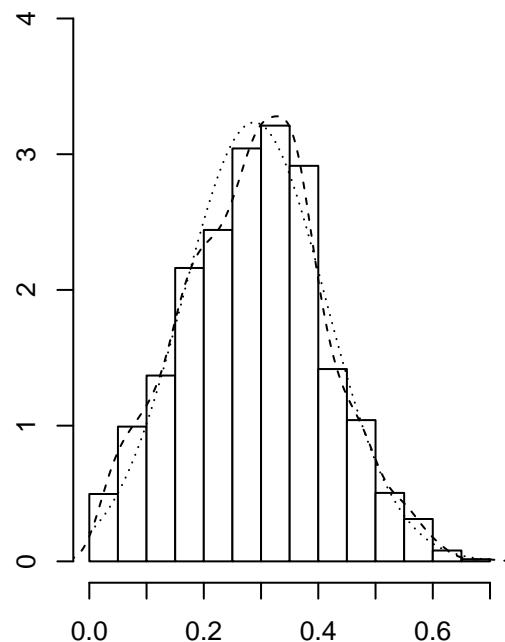
Histogram, Density, and Normal Fit



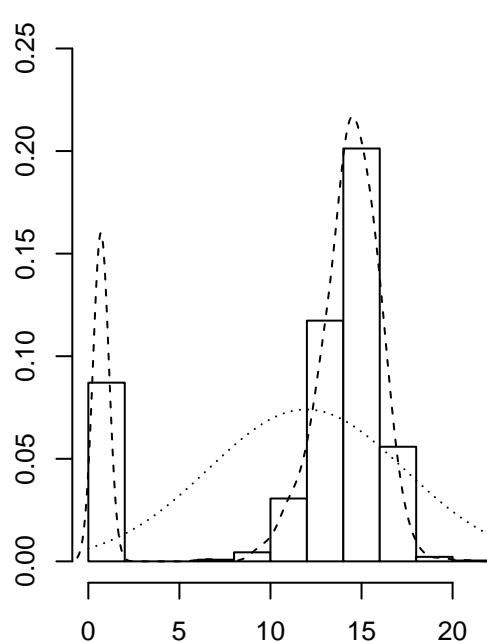
Histogram, Density, and Normal Fit



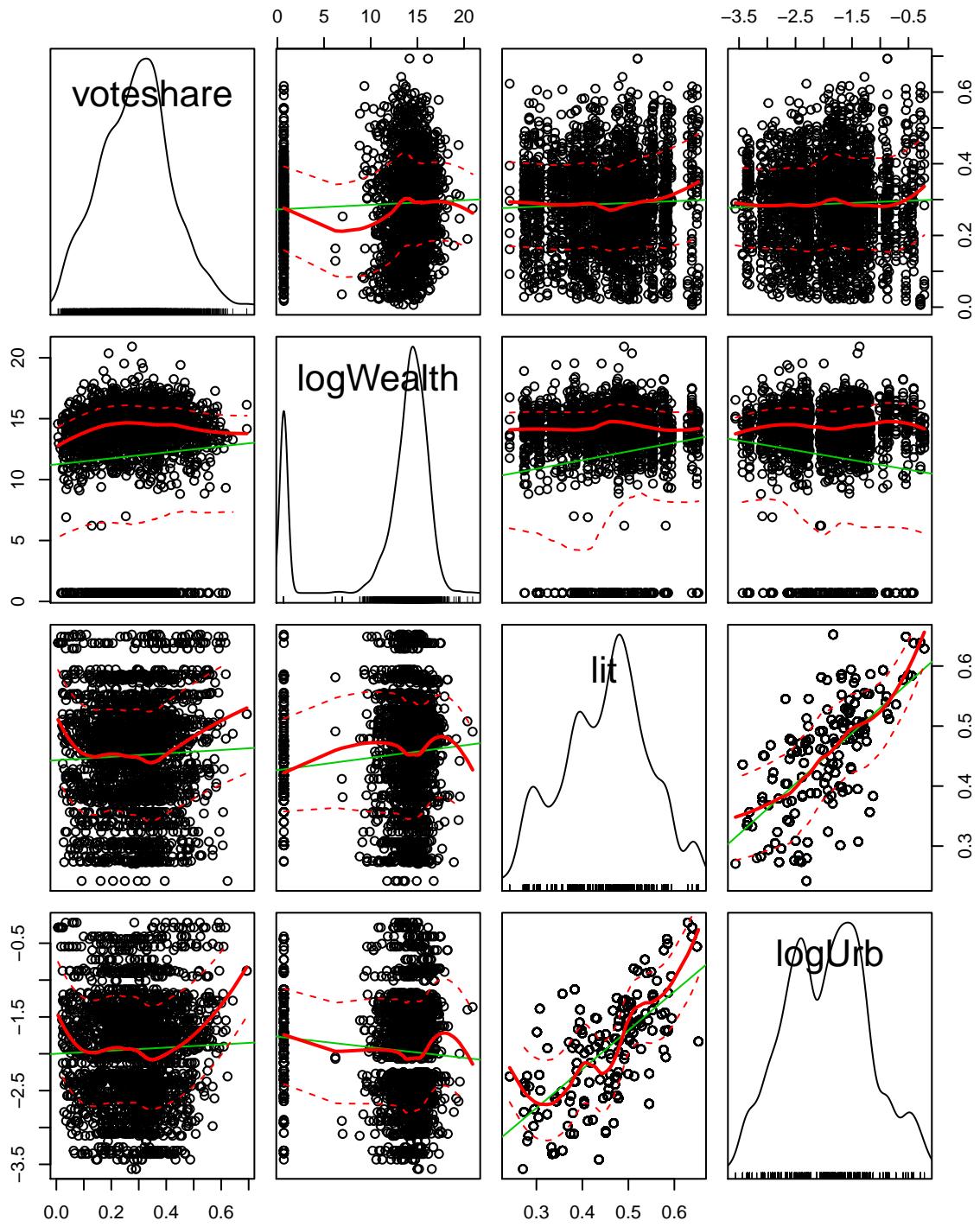
Histogram, Density, and Normal Fit



Histogram, Density, and Normal Fit



```
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) 0.27245805 0.00557433 48.8773 < 2.2e-16 ***
## logWealth   0.00129241 0.00041954  3.0806  0.002089 **
## ---
## 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.

```

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

```

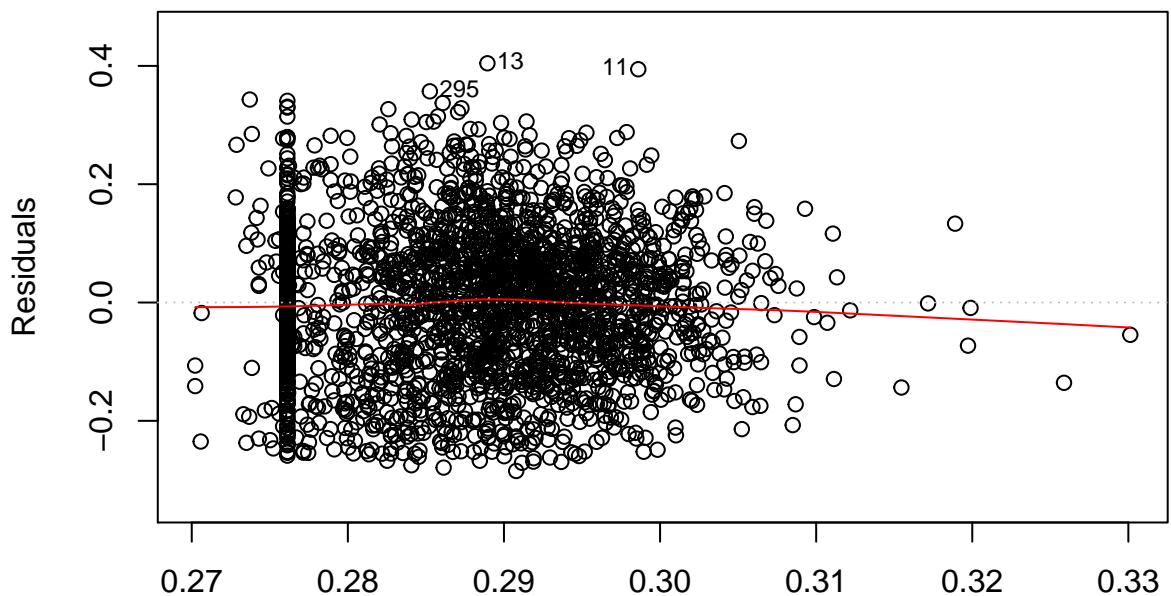
```

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.27796200 0.00611902 45.4259 < 2e-16 ***
## logWealth   -0.00282634 0.00230268 -1.2274 0.21978
## lwealthSquared 0.00025430 0.00013931  1.8254 0.06805 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

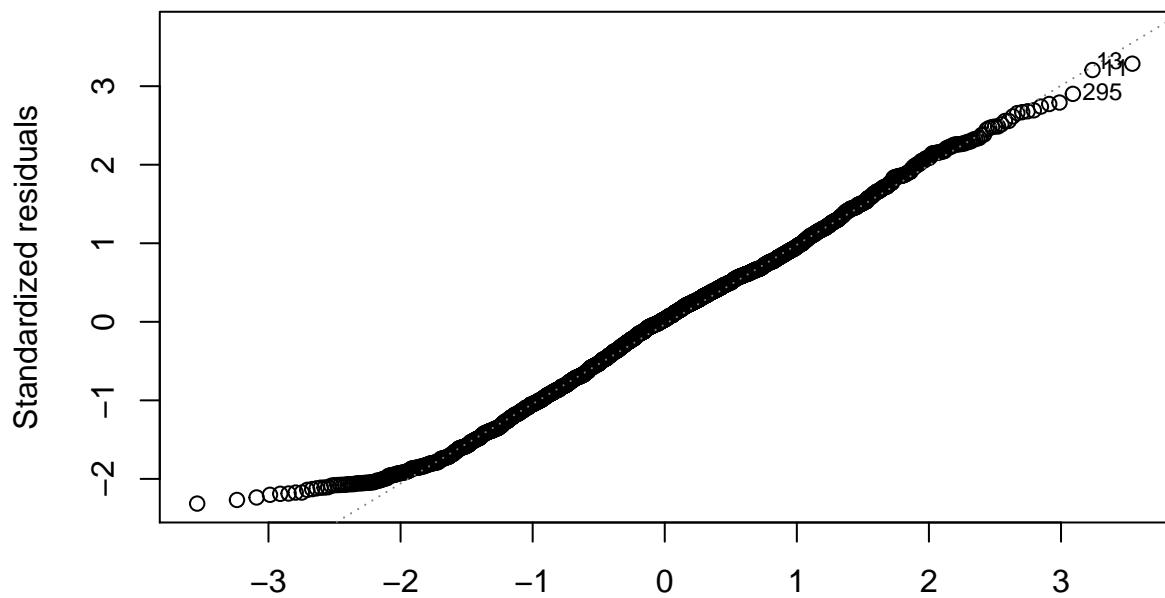
```

```
plot(model5.1.2)
```

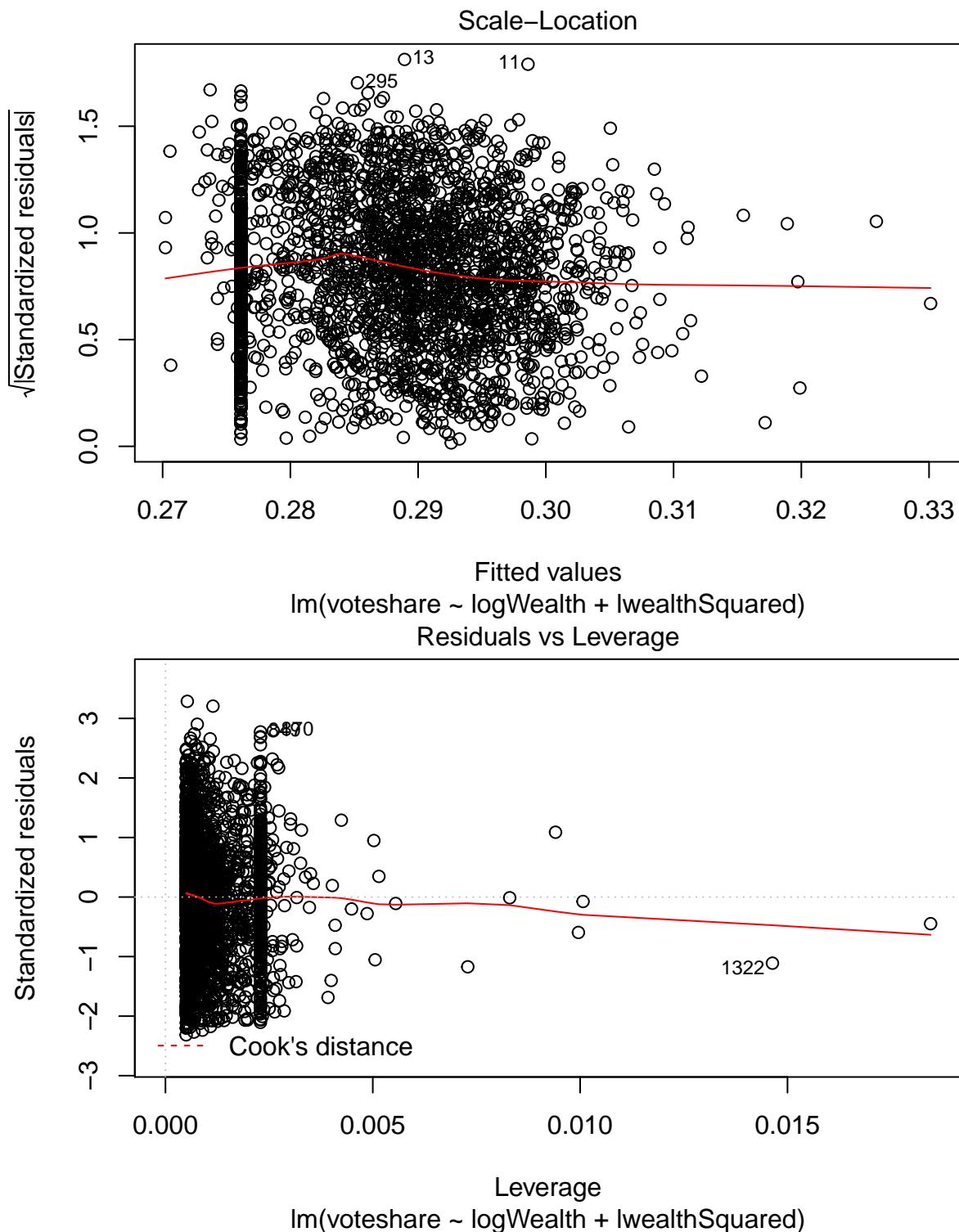
Residuals vs Fitted



Fitted values  
Im(voteshare ~ logWealth + lwealthSquared)  
Normal Q-Q



Theoretical Quantiles  
Im(voteshare ~ logWealth + lwealthSquared)



```
linearHypothesis(model5.1.2, c("lwealthSquared = 0"), vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```

## lwealthSquared = 0
##
## Model 1: restricted model
## Model 2: voteshare ~ logWealth + lwealthSquared
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1    2495
## 2    2494  1 3.3322 0.06805 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The results indicate a statistically significant relationship at the  $p = .04$  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
## 3.593724e-07

```

The null hypothesis that the  $\logWealth^2$  parameter is not statistically different than 0 is rejected with  $p = .04198$ .

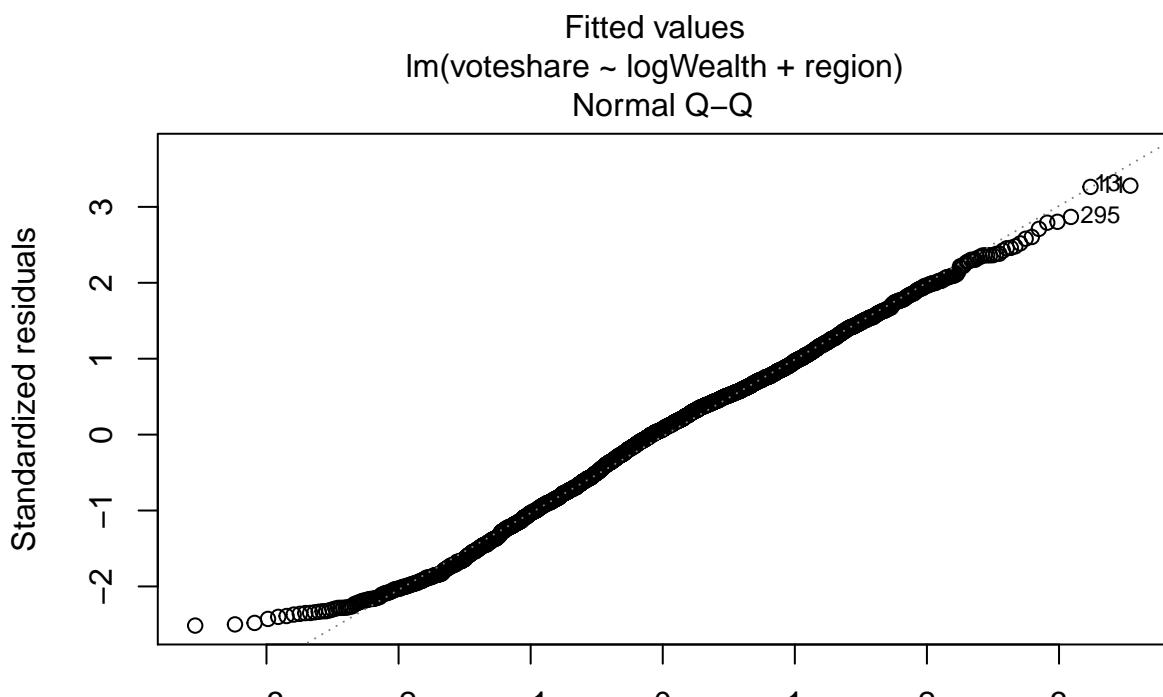
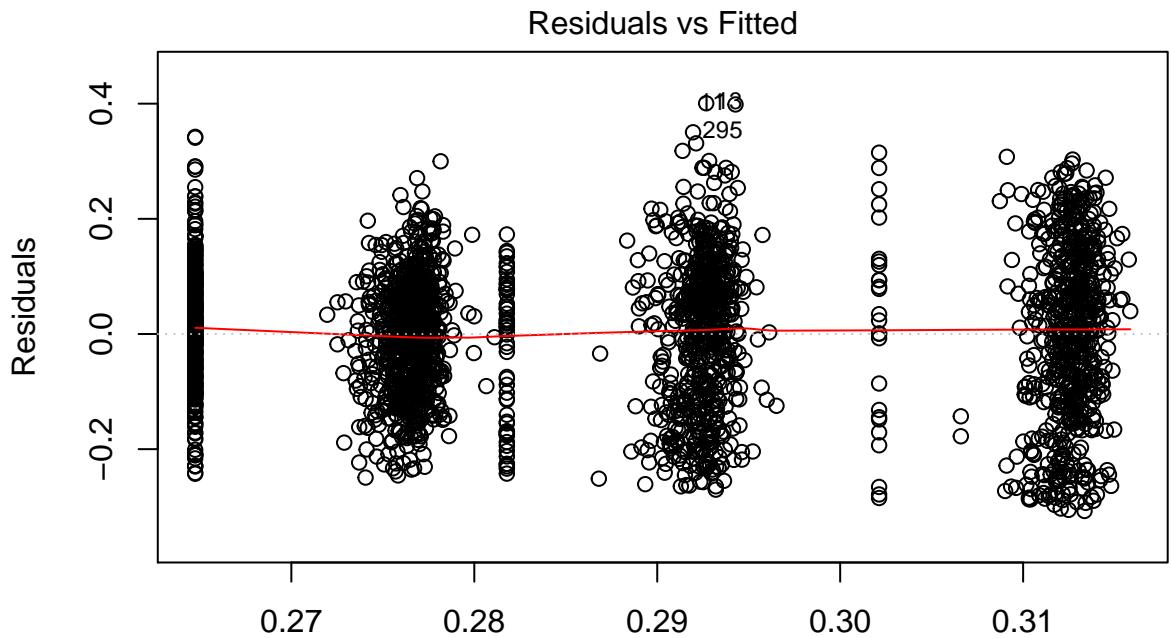
3. Another team member suggests that it is important to take into account the fact that different regions have different electoral contexts. In particular, the relationship between candidate wealth and electoral performance might be different across states. Modify your model and report your results.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

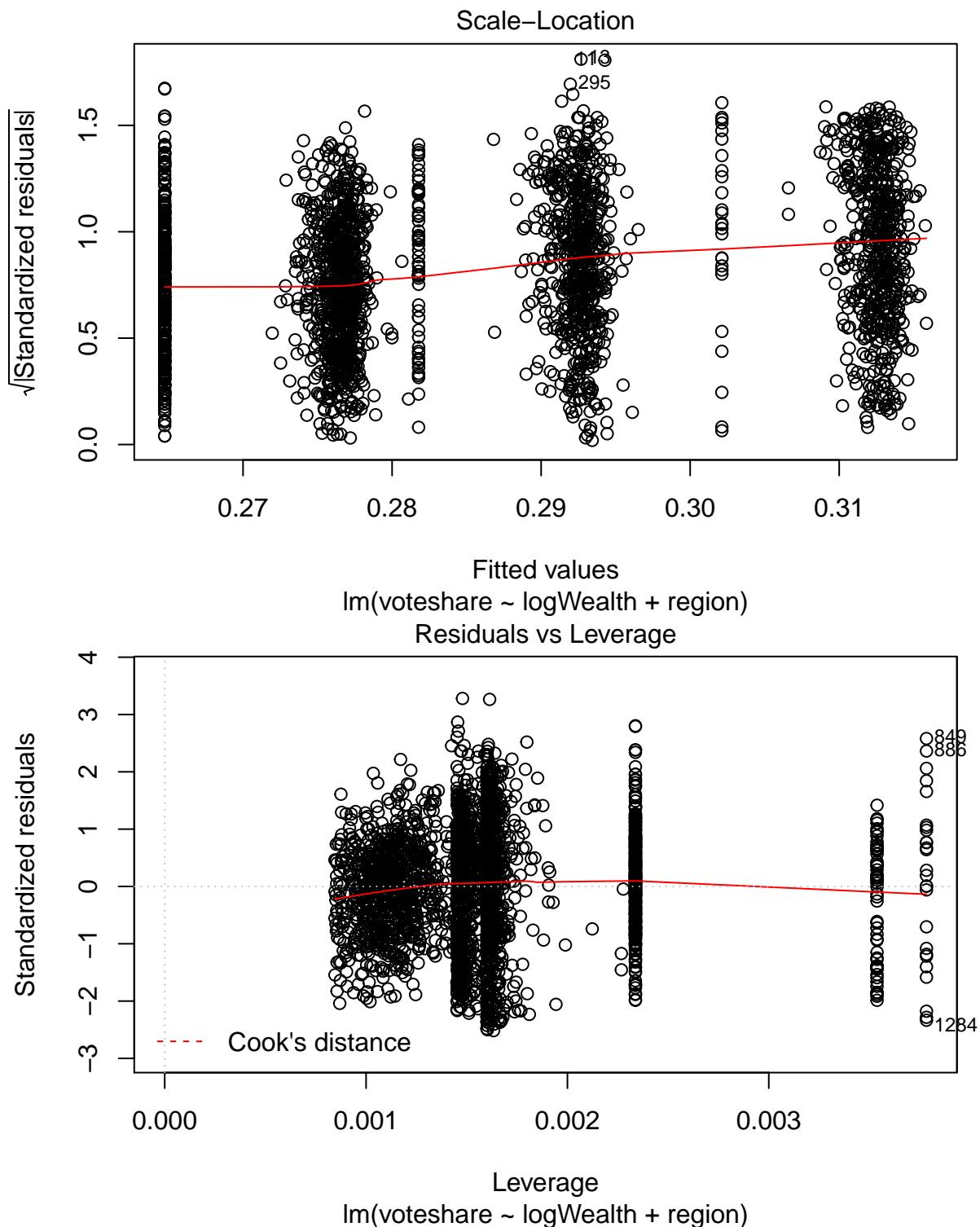
The *region* variable is already factored so we can use it in the model as is, with the base level as Region 1

```

model5.1.3 <- lm(voteshare ~ logWealth + region, data=df2)
plot(model5.1.3)

```





```
coefest(model5.1.3, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
```

```

## (Intercept) 0.26418229 0.00562395 46.9745 < 2.2e-16 ***
## logWealth 0.00080893 0.00042615 1.8982 0.057782 .
## regionRegion 2 0.01703745 0.00573420 2.9712 0.002995 **
## regionRegion 3 0.03738445 0.00679958 5.4981 4.231e-08 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

All the coefficients of the model are highly significant.

When wealth increases 1% voteshare increases by 1.2% in Region 1, assuming that voteshare is the margin of votes for the candidate. This effect increases in Region 2 by an additional 4% and in Region 3 by 6% over Region 1.

```
linearHypothesis(model5.1.3, c("regionRegion 2=0","regionRegion 3=0"), vcov=vcovHC)
```

```

## Linear hypothesis test
##
## Hypothesis:
## regionRegion 2 = 0
## regionRegion 3 = 0
##
## Model 1: restricted model
## Model 2: voteshare ~ logWealth + region
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1    2495
## 2    2493  2 16.705 6.215e-08 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Testing the null hypothesis that the *region* coefficients are no different from 0 results in the rejection of the null hypothesis with a highly significant result.

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 supported by data visualizations
- Give two reasons why the OLS model coefficients may be biased and/or not consistent, be specific.
- Propose (but do not actually implement) a plan for an IV approach to improve your forecasting model.

### Exploratory Data Analysis

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

```
str(retailSales)
```

```
## 'data.frame': 84672 obs. of 14 variables:
## $ Year : int 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 ...
## $ Product.type : Factor w/ 21 levels "Binoculars","Climbing Accessories",...: 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 ...
## $ Order.method.type: Factor w/ 7 levels "E-mail","Fax",...: 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 ...
```

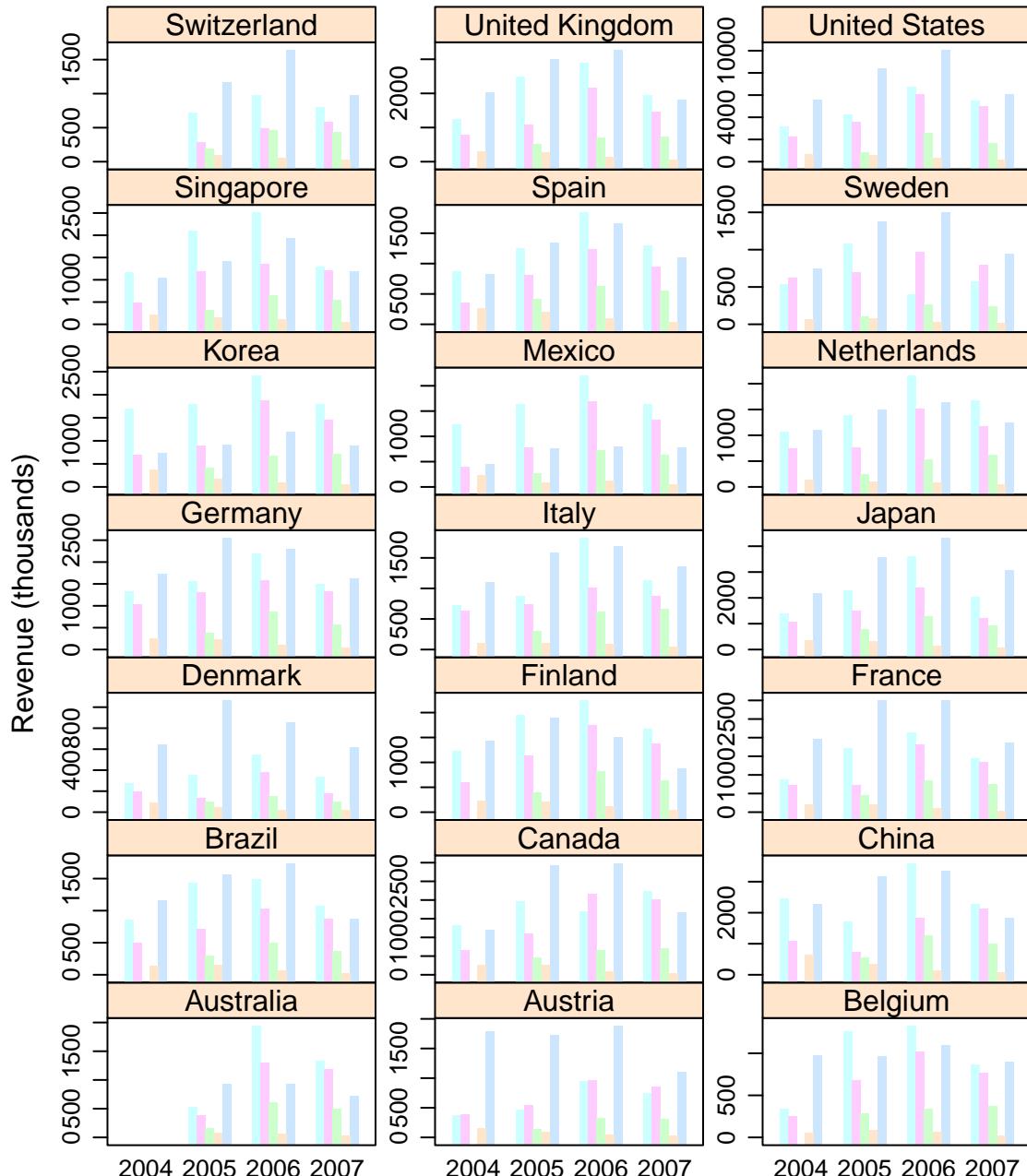
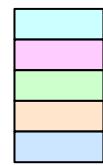
### Revenue and Projected Revenue

Revenue for 2007 is either declining from previous years or it is not a complete year's worth of data.

```
bc.gross_profit <- barchart(Revenue/1000.0 ~ as.factor(Year) | Retailer.country, groups=Product.line,
                             data=retailSales, scales=list(y="free"), layout=c(3,7),
                             auto.key=list(title="Product Lines", columns=1), ylab="Revenue (thousands)")
update(bc.gross_profit, border="transparent")
```

# Product Lines

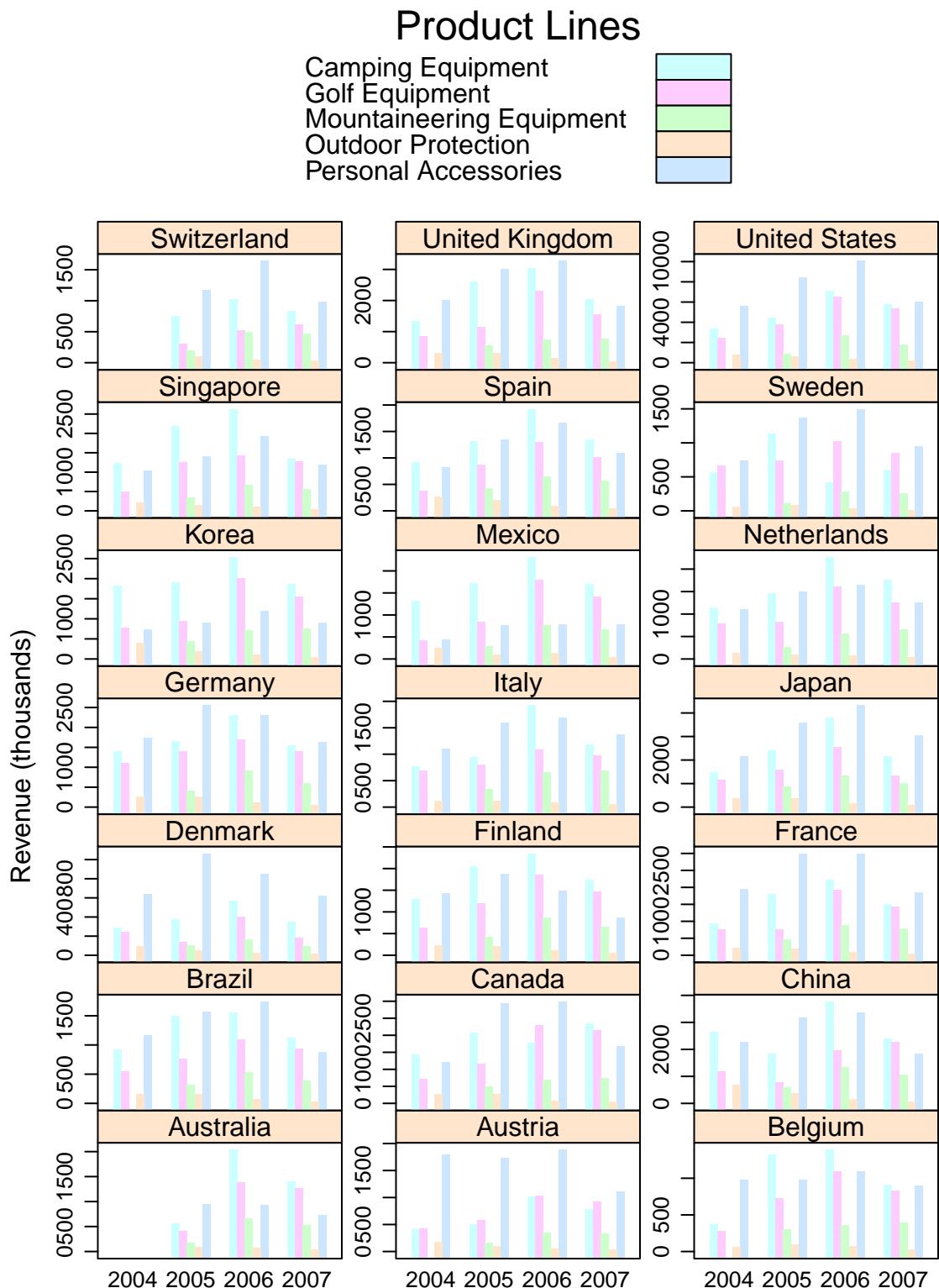
Camping Equipment  
 Golf Equipment  
 Mountaineering Equipment  
 Outdoor Protection  
 Personal Accessories



```

bc.gross_profit <- barchart(Planned.revenue/1000.0 ~ as.factor(Year) | Retailer.country, groups=Product
  data=retailSales, scales=list(y="free"), layout=c(3,7),
  auto.key=list(title="Product Lines", columns=1), ylab="Revenue (thousands)")
update(bc.gross_profit, border="transparent")

```



## **Gross Profit By Country**

The following two charts shows gross profit for each year broken into product lines. The first chart shows all countries on the same scale which shows that the United States is the largest retailer. The second chart is the same information presented such that each bar chart has an independent scale for gross profit. It's important to note that Switzerland and Australia have no sales data for 2004.

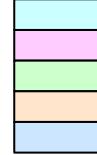
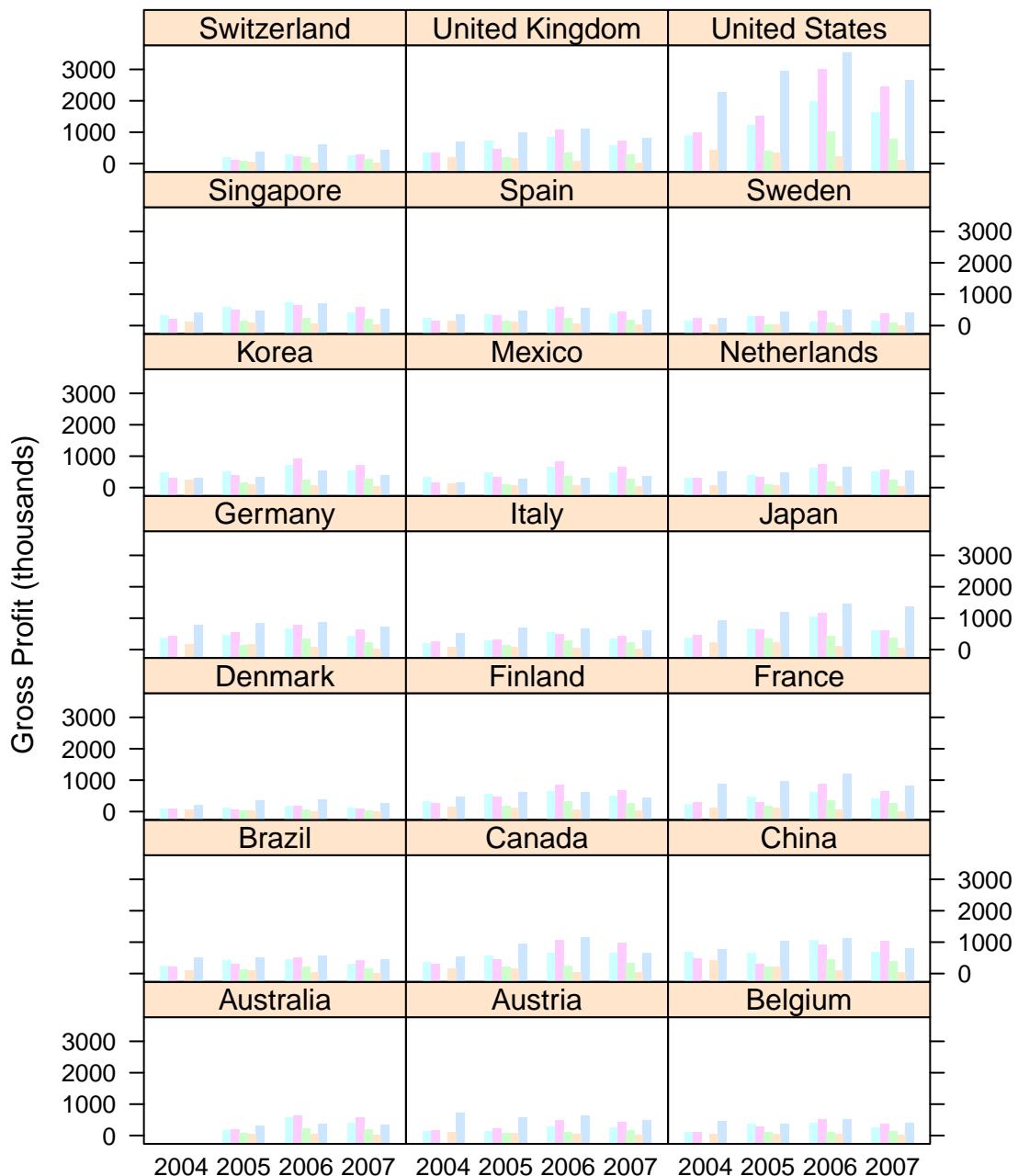
```

bc.gross_profit <- barchart(Gross.profit/1000.0 ~ as.factor(Year) | Retailer.country, groups=Product.line,
  data=retailSales, layout=c(3,7),
  auto.key=list(title="Product Lines", columns=1), ylab="Gross Profit (thousands)")
update(bc.gross_profit, border="transparent")

```

## Product Lines

Camping Equipment  
 Golf Equipment  
 Mountaineering Equipment  
 Outdoor Protection  
 Personal Accessories

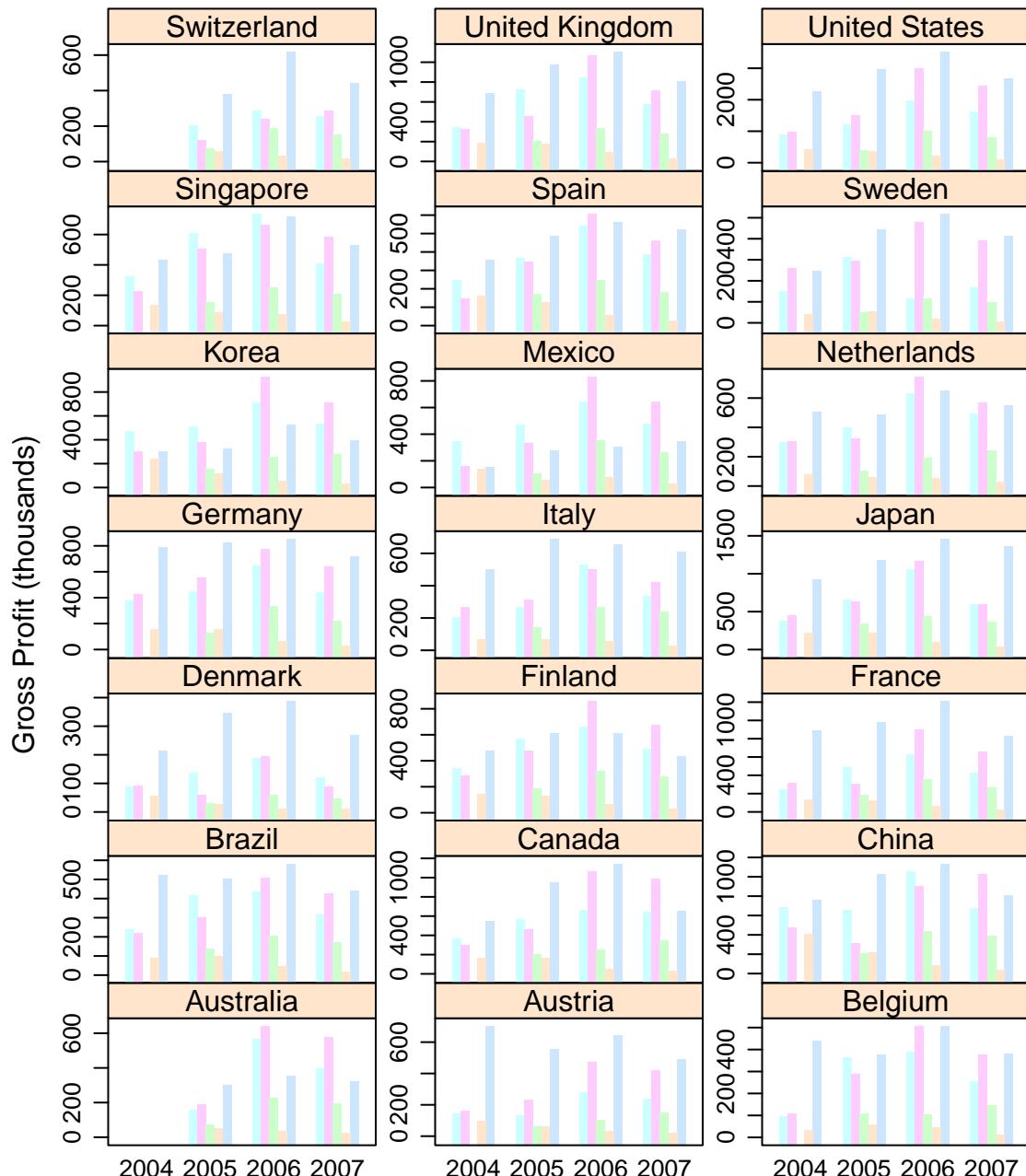
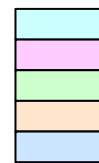



The Personal Accessory and Golf Equipment categories are responsible for the highest gross profits, with Camping Equipment following closely. The Outdoor Protection and Mountaineering Equipment product lines are not nearly as profitable.

```
bc.gross_profit <- barchart(Gross.profit/1000.0 ~ as.factor(Year) | Retailer.country, groups=Product.line  
    data=retailSales, scales=list(y="free"), layout=c(3,7),  
    auto.key=list(title="Product Lines", columns=1), ylab="Gross Profit (thousands)")  
update(bc.gross_profit, border="transparent")
```

# Product Lines

Camping Equipment  
 Golf Equipment  
 Mountaineering Equipment  
 Outdoor Protection  
 Personal Accessories

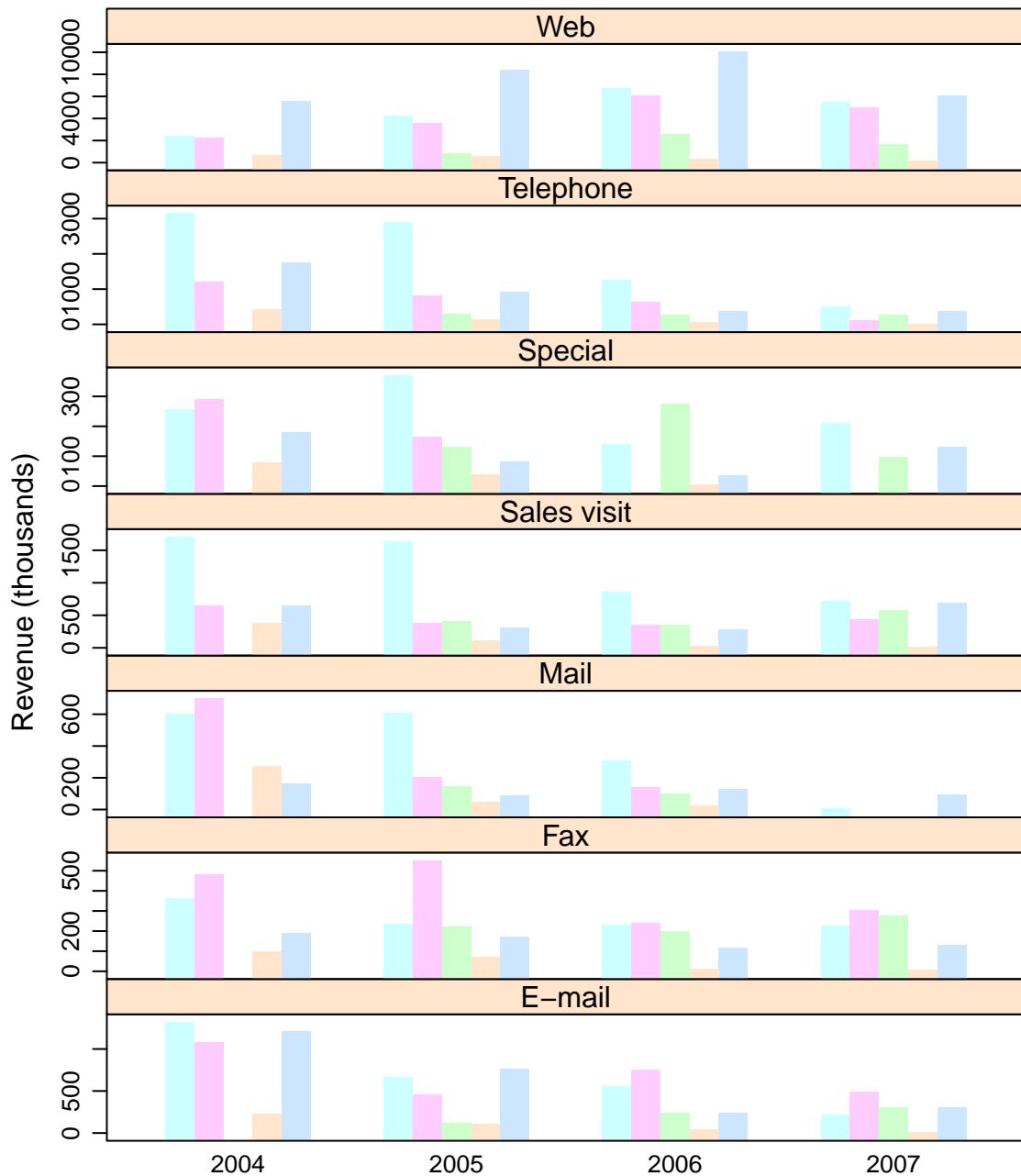


```
### Gross Profit By Order Method
```

```
bc.gross_profit <- barchart(Revenue/1000.0 ~ as.factor(Year) | Order.method.type, groups=Product.line,
  data=retailSales, scales=list(y="free"), layout=c(1,7),
  auto.key=list(title="Product Lines", columns=1), ylab="Revenue (thousands)")
update(bc.gross_profit, border="transparent")
```

## Product Lines

Camping Equipment  
 Golf Equipment  
 Mountaineering Equipment  
 Outdoor Protection  
 Personal Accessories

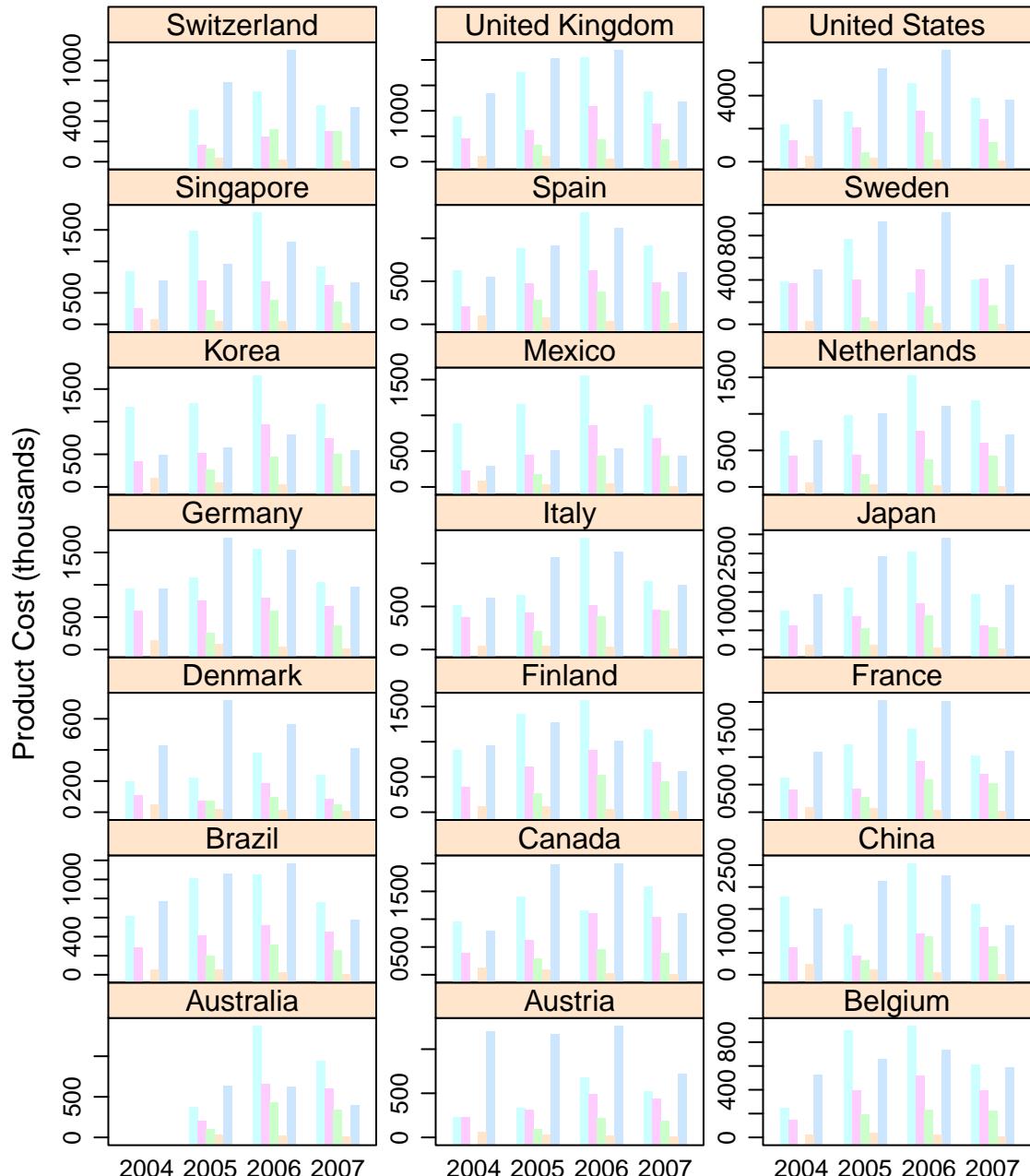
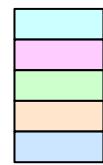


## Product Costs

```
bc.gross_profit <- barchart(Product.cost/1000.0 ~ as.factor(Year) | Retailer.country, groups=Product.line,
                             data=retailSales, scales=list(y="free"), layout=c(3,7),
                             auto.key=list(title="Product Lines", columns=1), ylab="Product Cost (thousands)")
update(bc.gross_profit, border="transparent")
```

# Product Lines

Camping Equipment  
 Golf Equipment  
 Mountaineering Equipment  
 Outdoor Protection  
 Personal Accessories

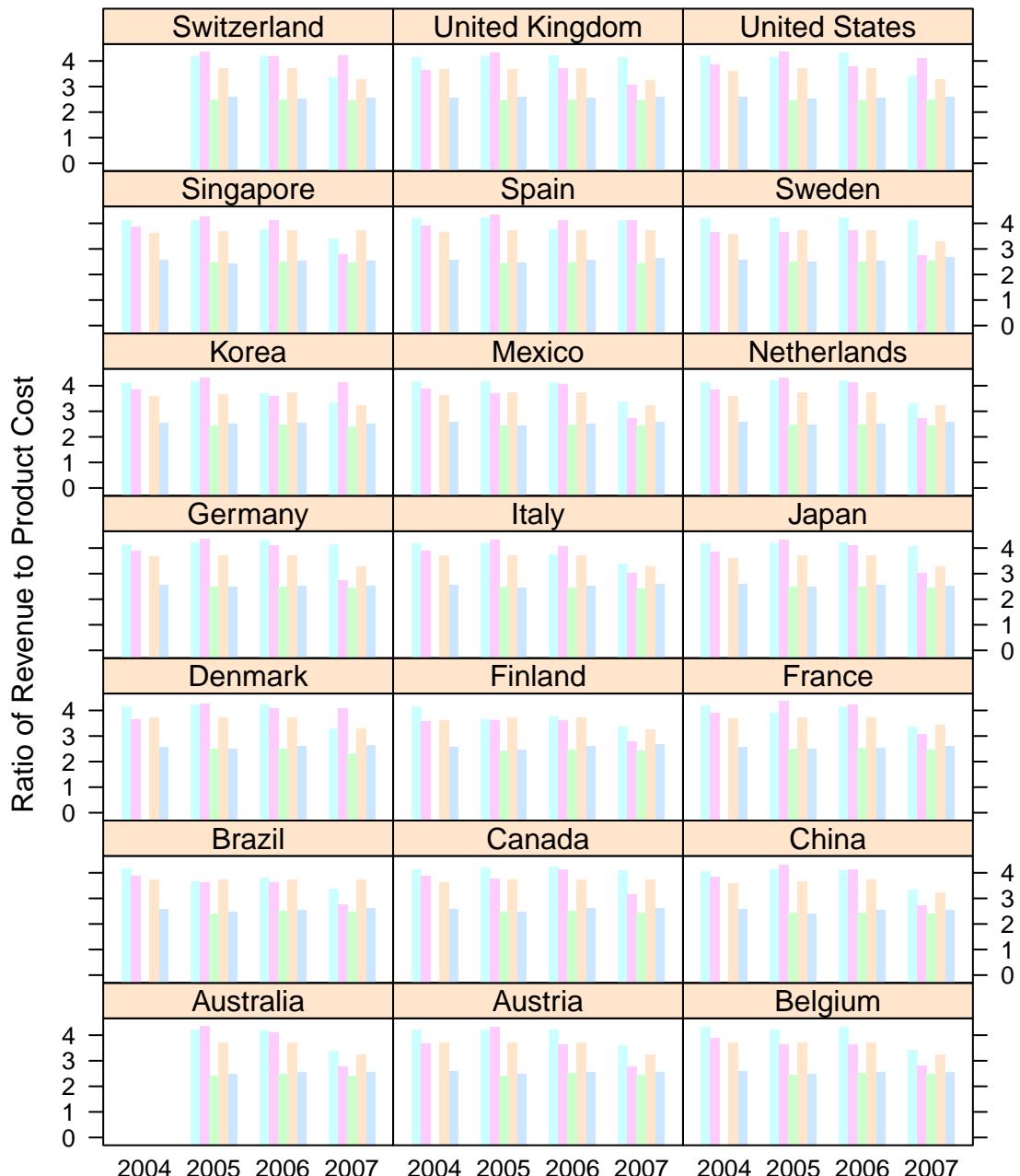


```
### Ratio of Revenue to Product Costs
```

```
bc.gross_profit <- barchart(Revenue/Product.cost ~ as.factor(Year) | Retailer.country, groups=Product.line,
                             data=retailSales, layout=c(3,7),
                             auto.key=list(title="Product Lines", columns=1), ylab="Ratio of Revenue to Product Cost")
update(bc.gross_profit, border="transparent")
```

# Product Lines

Camping Equipment  
 Golf Equipment  
 Mountaineering Equipment  
 Outdoor Protection  
 Personal Accessories



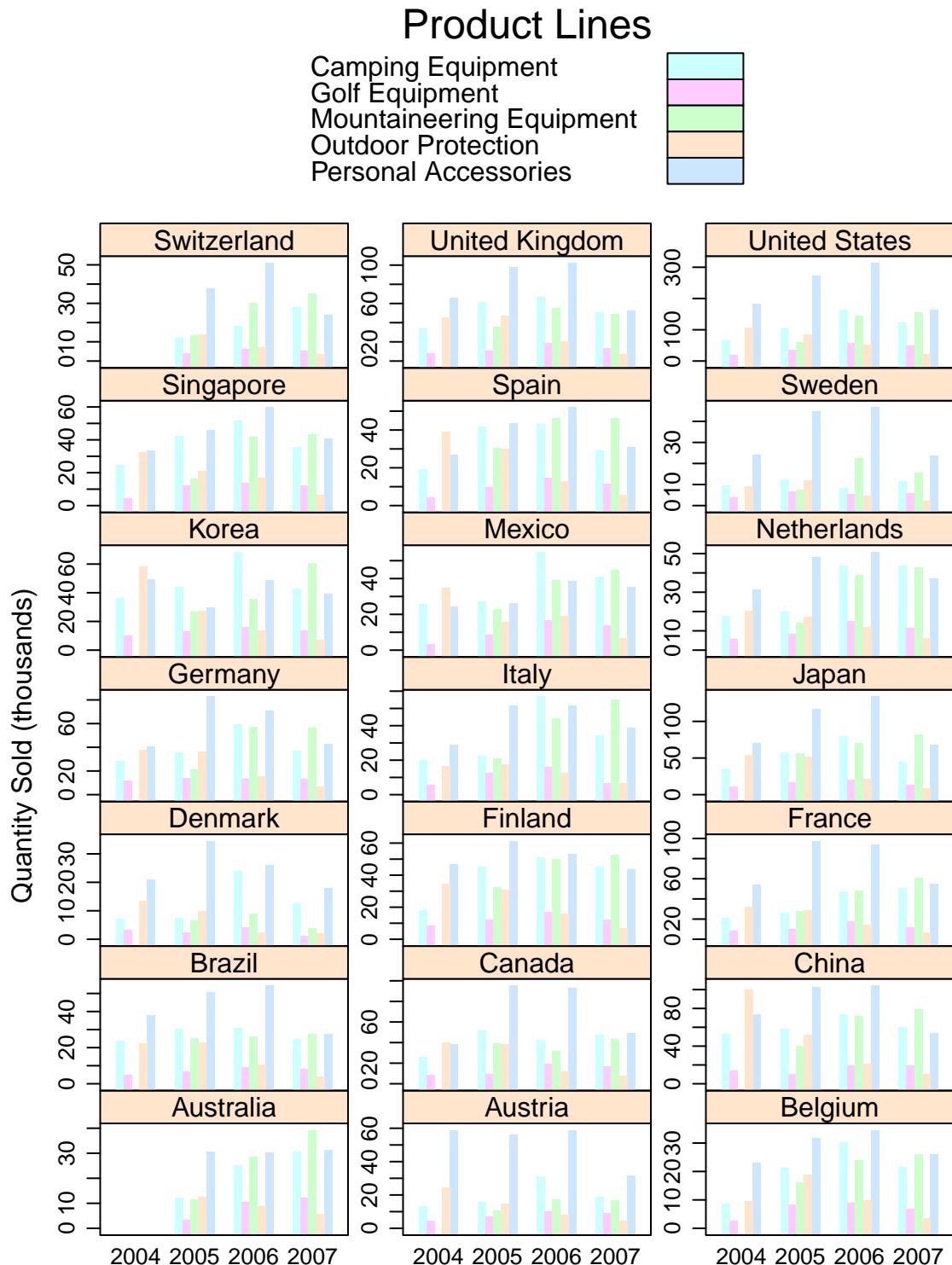
## Items Sold By Country

The leading product category by quantity is Personal Accessories, with Mountaineering Equipment and Camping Equipment following closely.

```

bc.quantity <- barchart(Quantity/1000.0 ~ as.factor(Year) | Retailer.country, groups=Product.line,
  data=retailSales, scales=list(y="free"), layout=c(3,7),
  auto.key=list(title="Product Lines", columns=1), ylab="Quantity Sold (thousands)")
update(bc.quantity, border="transparent")

```

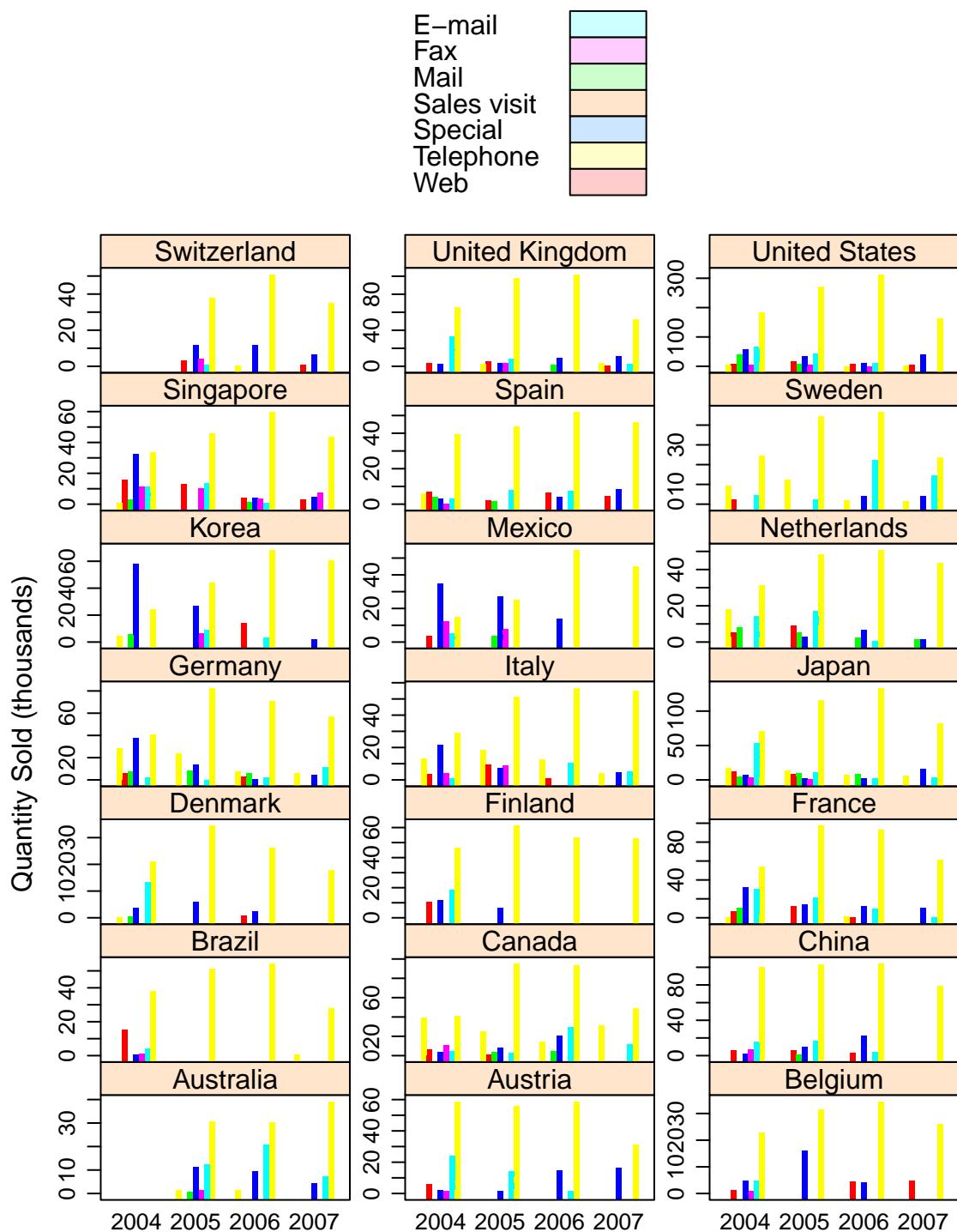


## Items Sold By Country and Order Method

In this chart is obvious that web sales is the dominant order method, so much so that most other sales have become nearly insignificant in most countries by 2006 and 2007.

```
bc.quantity <- barchart(Quantity/1000.0 ~ as.factor(Year) | Retailer.country, groups=Order.method.type,
  data=retailSales, scales=list(y="free"), layout=c(3,7),
  auto.key=list(title="Order Methods", columns=1), ylab="Quantity Sold (thousands)",
  col=c('yellow','red','green','blue','magenta','cyan'))
update(bc.quantity, border="transparent")
```

## Order Methods

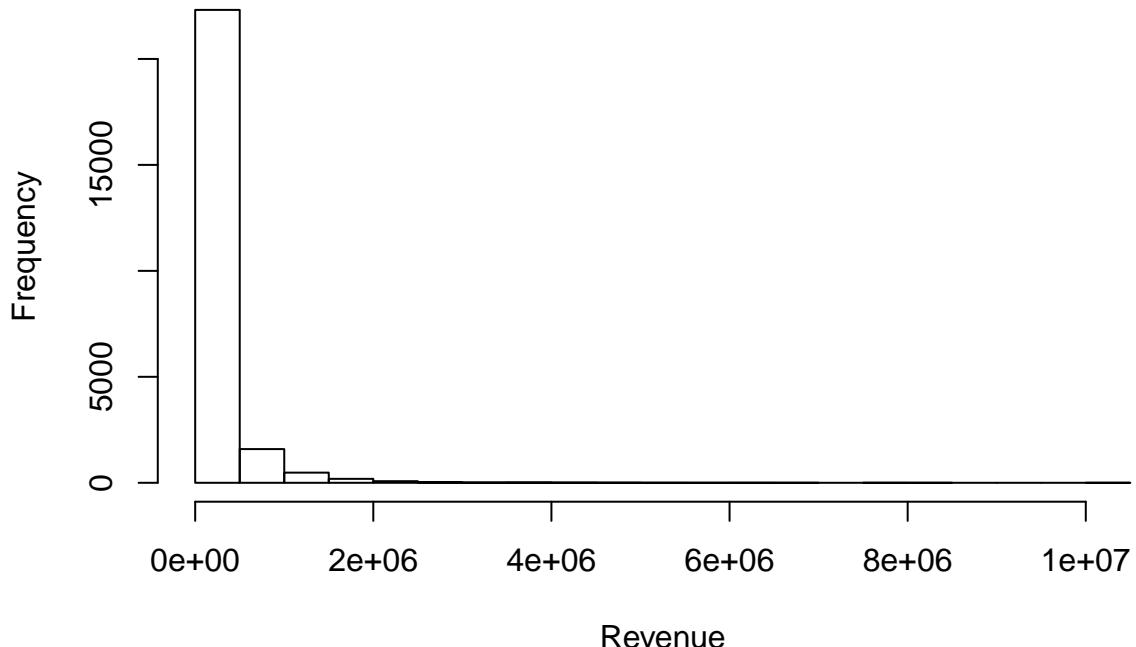


## Variable Checking

The *Revenue* variable is highly right skewed but a log transformation helps with this. Also, it appears that the revenue for 2007 doesn't contain a full year of data judging from the overall numbers. Alternatively the revenue is in decline for 2007. However it might be a good idea to evaluate 2007 as if it is missing data from the second half of the year.

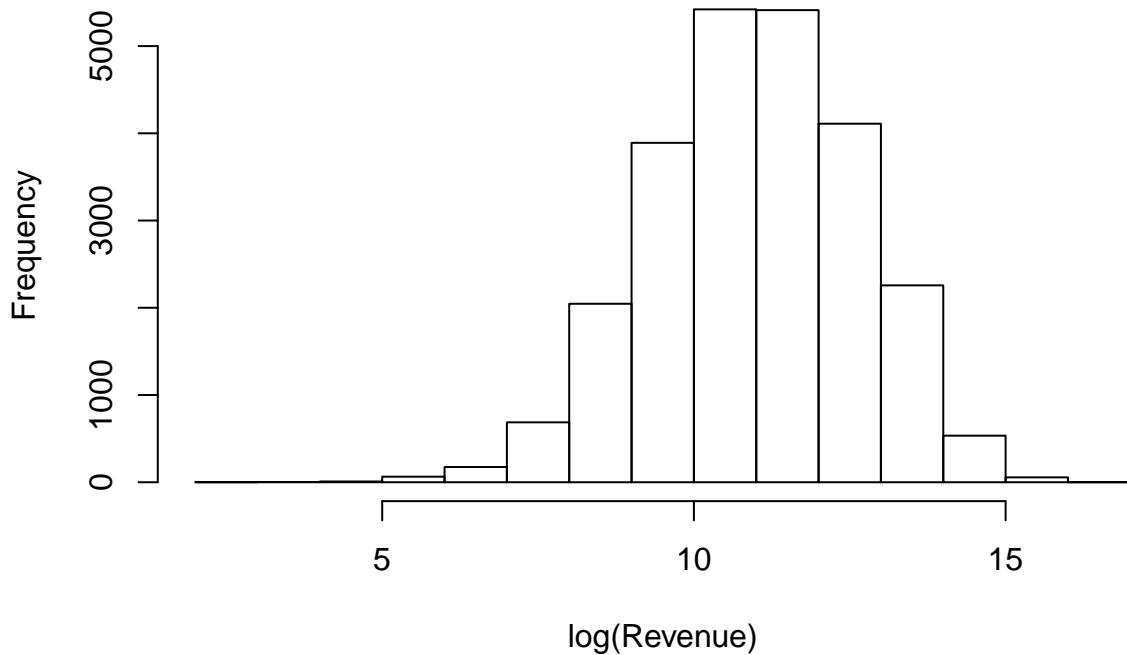
```
hist(retailSales$Revenue, main='Revenue Histogram', xlab='Revenue')
```

**Revenue Histogram**



```
hist(log(retailSales$Revenue), main='Log(Revenue) Histogram', xlab='log(Revenue)')
```

## Log(Revenue) Histogram



```
summary(retailSales$Revenue)
```

```
##      Min.    1st Qu.     Median      Mean    3rd Qu.      Max.    NA's
##      0       18580     59870    189400   190200  10050000    59929
```

```
summary(retailSales$Planned.revenue)
```

```
##      Min.    1st Qu.     Median      Mean    3rd Qu.      Max.    NA's
##      16      19560     63910    198800   204000  10050000    59929
```

As an investigation compare an “adjusted” 2007 revenue value that is twice the given amount and visualize as a graph.

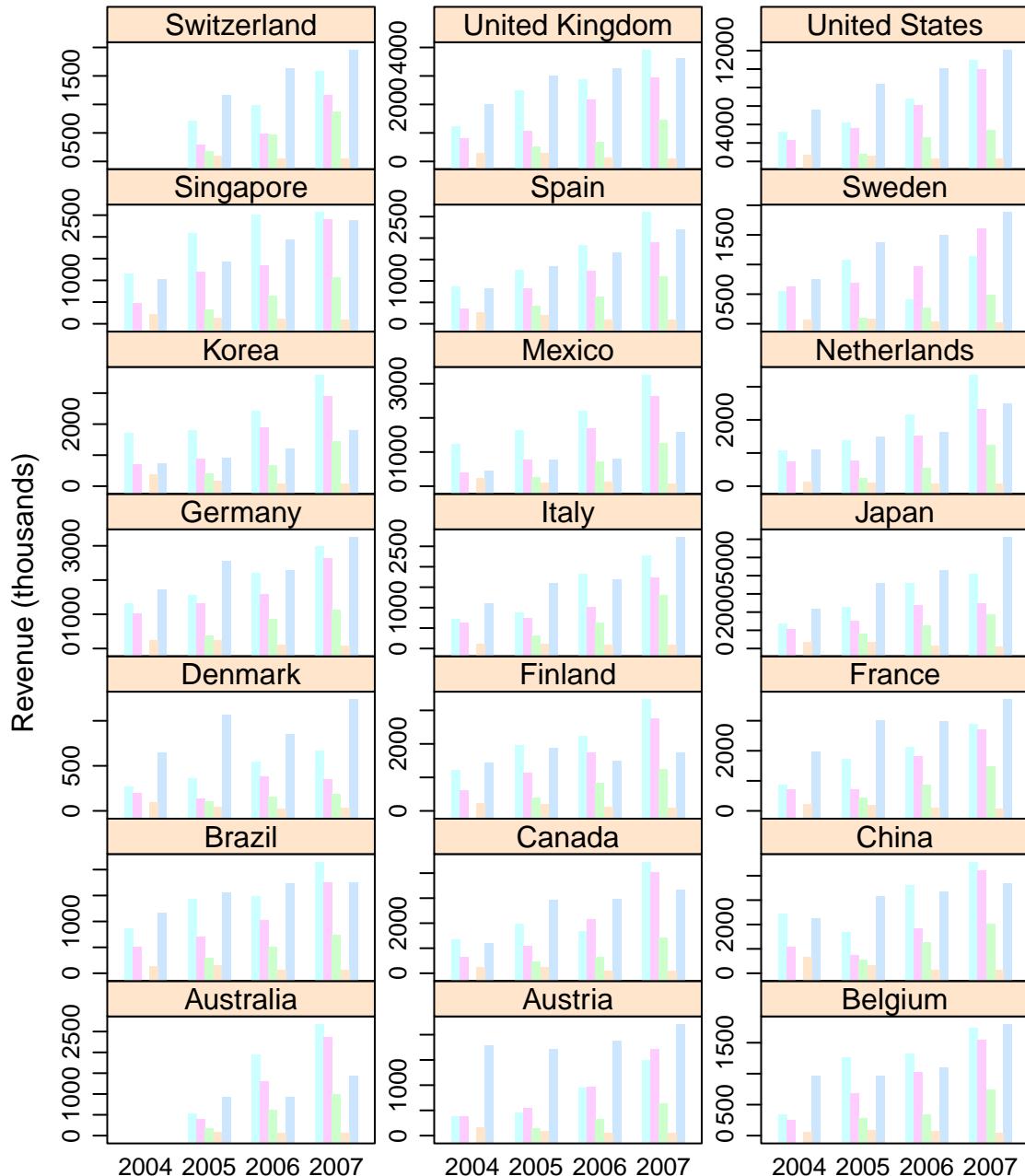
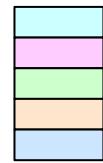
```
# Investigate revenue for 2007
rs.test <- retailSales
rs.test$Revenue <- ifelse(rs.test$Year==2007, 2*rs.test$Revenue, rs.test$Revenue)
```

The graph would indicate that we are probably missing the last half of 2007 data. We should keep this in mind when evaluating models.

```
bc.adj_revenue <- barchart(Revenue/1000.0 ~ as.factor(Year) | Retailer.country, groups=Product.line,
                           data=rs.test, scales=list(y="free"), layout=c(3,7),
                           auto.key=list(title="Product Lines", columns=1), ylab="Revenue (thousands)")
update(bc.adj_revenue, border="transparent")
```

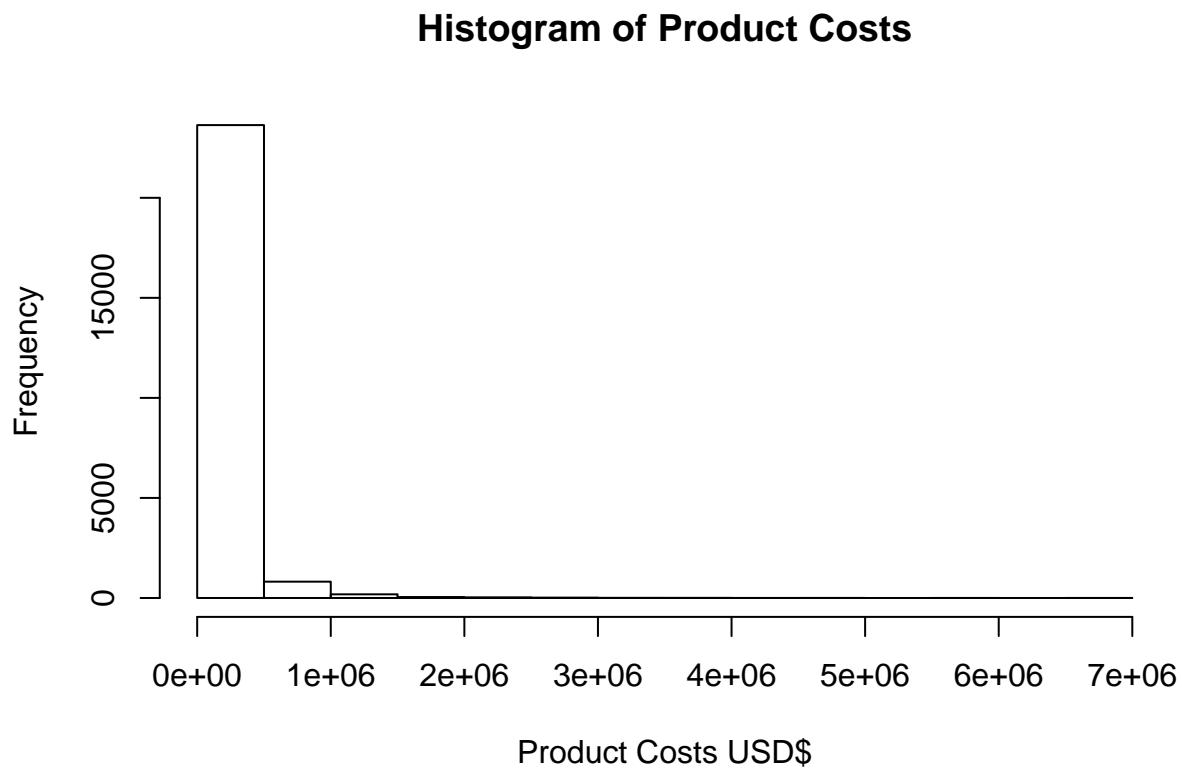
# Product Lines

Camping Equipment  
Golf Equipment  
Mountaineering Equipment  
Outdoor Protection  
Personal Accessories

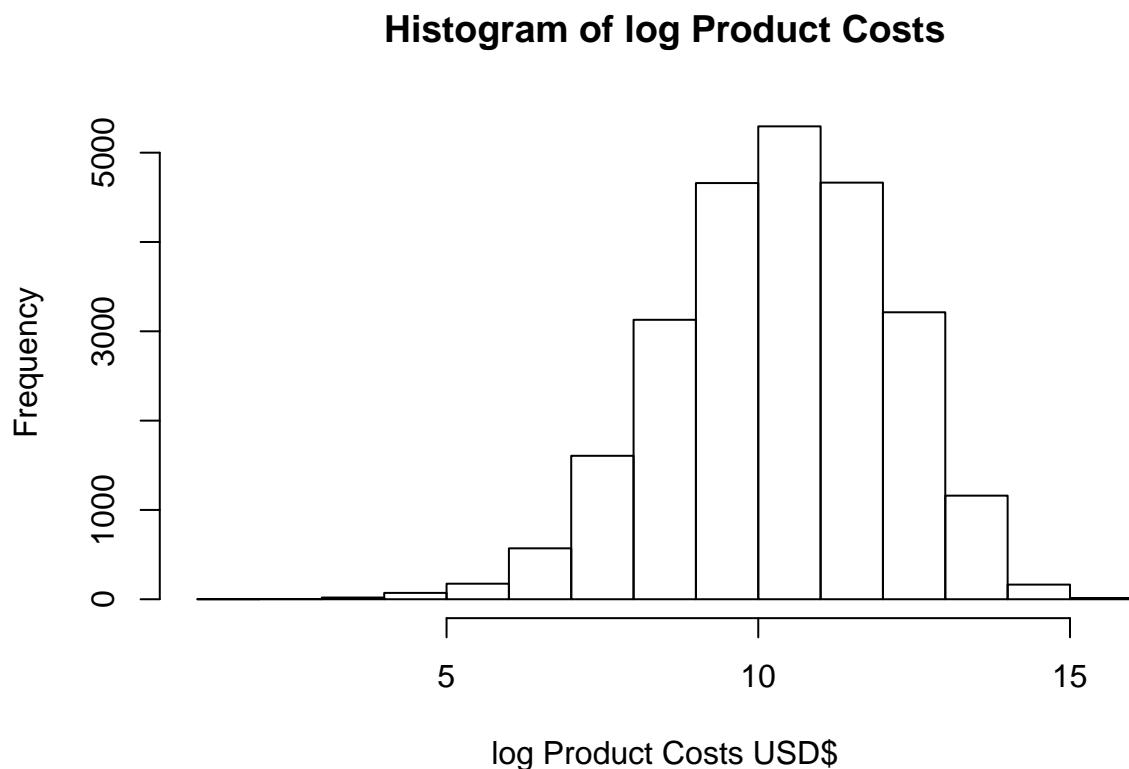


Product Costs histogram shows that we should use the log of product costs to deal with the significant skew in the untransformed variable.

```
hist( retailSales$Product.cost, main='Histogram of Product Costs', xlab='Product Costs USD$' )
```



```
hist(log(retailSales$Product.cost), main='Histogram of log Product Costs', xlab='log Product Costs USD$')
```



## Model Building

### Simple Model - Revenue by Country

```
# First aggregate revenue by country across all attributes
revenues <- group_by(retailSales, Retailer.country, Year)
revenue.by_country <- summarise(revenues, revenues = sum(Revenue, na.rm=TRUE))
# split into train and test
revenue.by_country.train <- subset(revenue.by_country, Year<2006)
revenue.by_country.test <- subset(revenue.by_country, Year>2005)
# create a simple OLS model for log(revenue) ~ country
model.rev.country <- lm(log1p(revenues) ~ Retailer.country,
                         data=revenue.by_country.train )
summary(model.rev.country)

##
## Call:
## lm(formula = log1p(revenues) ~ Retailer.country, data = revenue.by_country.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -8.5423 -0.0922  0.0000  0.0922  8.5423 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 8.5292    2.6375   3.234  0.00398 ** 
## Retailer.countryAustria 8.7641    3.7299   2.350  0.02865 *  
## Retailer.countryBelgium 8.3820    3.7299   2.247  0.03550 *  
## Retailer.countryBrazil 8.7402    3.7299   2.343  0.02904 *  
## Retailer.countryCanada 9.4890    3.7299   2.544  0.01889 *  
## Retailer.countryChina 9.5844    3.7299   2.570  0.01787 *  
## Retailer.countryDenmark 7.9929    3.7299   2.143  0.04399 *  
## Retailer.countryFinland 9.0168    3.7299   2.417  0.02481 *  
## Retailer.countryFrance 9.5290    3.7299   2.555  0.01845 *  
## Retailer.countryGermany 9.4352    3.7299   2.530  0.01949 *  
## Retailer.countryItaly 8.9353    3.7299   2.396  0.02599 *  
## Retailer.countryJapan 9.6794    3.7299   2.595  0.01690 *  
## Retailer.countryKorea 9.1213    3.7299   2.445  0.02337 *  
## Retailer.countryMexico 8.8223    3.7299   2.365  0.02772 *  
## Retailer.countryNetherlands 8.9663    3.7299   2.404  0.02554 *  
## Retailer.countrySingapore 9.0852    3.7299   2.436  0.02386 *  
## Retailer.countrySpain 8.7926    3.7299   2.357  0.02819 *  
## Retailer.countrySweden 8.2212    3.7299   2.204  0.03881 *  
## Retailer.countrySwitzerland 0.0131    3.7299   0.004  0.99723  
## Retailer.countryUnited Kingdom 9.4391    3.7299   2.531  0.01945 *  
## Retailer.countryUnited States 10.5246   3.7299   2.822  0.01022 *  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.73 on 21 degrees of freedom
## Multiple R-squared:  0.515, Adjusted R-squared:  0.05308 
## F-statistic: 1.115 on 20 and 21 DF, p-value: 0.4025
```

```

# predict the outcome of the test data set
model.rev.country.pred <- predict(model.rev.country, revenue.by_country.test)
# measure against test data
model.rev.country.lmse <- sqrt(sum((model.rev.country.pred -
                                         log1p(revenue.by_country.test$revenues))^2))
model.rev.country.lmse

## [1] 17.98069

```

The model summary indicates single coefficient testing indicates that all coefficients are significant except Switzerland. This is expected since Switzerland does not have 2004 sales data. However, we notice that Australia doesn't have 2004 data either and it's used as the basis of the country factor variable. The  $R^2 = 0.05308$  indicates that this model doesn't account for very much of the variation in revenue. The model is not significant with an F-statistic of .49.

Let's evaluate the model against the 2006-2007 test data set, keeping in mind that 2007 test data may not be complete.

Evaluate against 2006 only:

```

# predict the outcome of the test data set for only 2006
model.rev.country.pred <- predict(model.rev.country, subset(revenue.by_country.test, Year==2006))
# measure against test data
model.rev.country.lmse <- sqrt(sum((model.rev.country.pred -
                                         log1p(subset(revenue.by_country.test, Year==2006)$revenues))^2))
model.rev.country.lmse

## [1] 12.93839

```

### Simple Model 2: Adding Year to Revenue By Country

```

# create a simple OLS model for log(revenue) ~ country
model.rev.country_year <- lm(log1p(revenues) ~ Retailer.country + Year,
                               data=revenue.by_country.train )
summary(model.rev.country_year)

##
## Call:
## lm(formula = log1p(revenues) ~ Retailer.country + Year, data = revenue.by_country.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -7.6236 -0.8373  0.0000  0.8373  7.6236 
##
## Coefficients:
## (Intercept)          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -3674.5140   2216.2628  -1.658  0.11292  
## Retailer.countryAustria    8.7641    3.5827   2.446  0.02381 *  
## Retailer.countryBelgium    8.3820    3.5827   2.340  0.02979 *  
## Retailer.countryBrazil    8.7402    3.5827   2.440  0.02415 *  
## Retailer.countryCanada    9.4890    3.5827   2.649  0.01541 *  

```

```

## Retailer.countryChina          9.5844   3.5827   2.675  0.01455 *
## Retailer.countryDenmark       7.9929   3.5827   2.231  0.03729 *
## Retailer.countryFinland        9.0168   3.5827   2.517  0.02049 *
## Retailer.countryFrance         9.5290   3.5827   2.660  0.01504 *
## Retailer.countryGermany        9.4352   3.5827   2.634  0.01593 *
## Retailer.countryItaly           8.9353   3.5827   2.494  0.02151 *
## Retailer.countryJapan           9.6794   3.5827   2.702  0.01373 *
## Retailer.countryKorea           9.1213   3.5827   2.546  0.01924 *
## Retailer.countryMexico          8.8223   3.5827   2.462  0.02300 *
## Retailer.countryNetherlands     8.9663   3.5827   2.503  0.02111 *
## Retailer.countrySingapore        9.0852   3.5827   2.536  0.01966 *
## Retailer.countrySpain            8.7926   3.5827   2.454  0.02341 *
## Retailer.countrySweden           8.2212   3.5827   2.295  0.03271 *
## Retailer.countrySwitzerland      0.0131   3.5827   0.004  0.99712
## Retailer.countryUnited Kingdom    9.4391   3.5827   2.635  0.01589 *
## Retailer.countryUnited States     10.5246  3.5827   2.938  0.00814 **
## Year                            1.8374   1.1056   1.662  0.11214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.583 on 20 degrees of freedom
## Multiple R-squared:  0.5738, Adjusted R-squared:  0.1264
## F-statistic: 1.282 on 21 and 20 DF,  p-value: 0.2907

# predict the outcome of the test data set
model.rev.country_year.pred <- predict(model.rev.country_year, revenue.by_country.test)
# measure against test data
model.rev.country_year.lmse <- sqrt(sum((model.rev.country_year.pred -
                                             log1p(revenue.by_country.test$revenues))^2))
model.rev.country_year.lmse
```

## [1] 24.8136

The  $R^2 = 0.1264$  is better but the F-statistic shows that the model is not significant,  $p = .291$  and the Log Mean Square Error increased to 12.8136

```

# predict the outcome of the test data set for only 2006
model.rev.country_year.pred <- predict(model.rev.country_year, subset(revenue.by_country.test, Year==2006))
# measure against test data
model.rev.country_year.lmse <- sqrt(sum((model.rev.country_year.pred -
                                             log1p(subset(revenue.by_country.test, Year==2006)$revenues))^2))
model.rev.country_year.lmse
```

## [1] 13.75788

Prediction 2006 data drops the LMSE back down to 13.75788 but still larger than with country only.

### Model 3: Adding in Product Costs

```

# First aggregate revenue by country across all attributes
revenue.by_prod <-
  summarise(group_by(retailSales, Product.line, Retailer.country, Year),
    revenues = sum(Revenue, na.rm=TRUE),
    prodCosts = sum(Product.cost, na.rm=TRUE),
    plannedRev = sum(Planned.revenue, na.rm=TRUE))

# transform the revenue and product cost variables
revenue.by_prod$revenues <- log1p(revenue.by_prod$revenues)
revenue.by_prod$prodCosts <- log1p(revenue.by_prod$prodCosts)

# split into train and test
revenue.by_prod.train <- subset(revenue.by_prod, Year<2006)
revenue.by_prod.test <- subset(revenue.by_prod, Year>2005)

# create a simple OLS model for log(revenue) ~ Product.cost
model.rev.prod <- lm(revenues ~ prodCosts + Product.line + Retailer.country + Year,
                      data=revenue.by_prod.train)
summary(model.rev.prod)

## 
## Call:
## lm(formula = revenues ~ prodCosts + Product.line + Retailer.country +
##     Year, data = revenue.by_prod.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42357 -0.01758 -0.00394  0.01889  0.13806
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)             -64.447608  17.693848 -3.642
## prodCosts                  1.034109  0.001110 931.614
## Product.lineGolf Equipment  0.204886  0.012064 16.983
## Product.lineMountaineering Equipment  0.096320  0.014899  6.465
## Product.lineOutdoor Protection  0.560595  0.012427 45.109
## Product.linePersonal Accessories  0.062780  0.012016  5.225
## Retailer.countryAustria  0.046672  0.025495  1.831
## Retailer.countryBelgium  0.055614  0.025426  2.187
## Retailer.countryBrazil  0.042730  0.025516  1.675
## Retailer.countryCanada  0.018250  0.025733  0.709
## Retailer.countryChina  0.007813  0.025768  0.303
## Retailer.countryDenmark  0.068253  0.025305  2.697
## Retailer.countryFinland  0.025081  0.025613  0.979
## Retailer.countryFrance  0.014083  0.025740  0.547
## Retailer.countryGermany  0.015021  0.025720  0.584
## Retailer.countryItaly  0.034149  0.025584  1.335
## Retailer.countryJapan  0.010057  0.025791  0.390
## Retailer.countryKorea  0.018897  0.025656  0.737
## Retailer.countryMexico  0.031298  0.025570  1.224
## Retailer.countryNetherlands  0.030955  0.025591  1.210
## Retailer.countrySingapore  0.022379  0.025628  0.873
## Retailer.countrySpain  0.037477  0.025552  1.467

```

```

## Retailer.countrySweden          0.055538  0.025374  2.189
## Retailer.countrySwitzerland     0.001034  0.024624  0.042
## Retailer.countryUnited Kingdom  0.011188  0.025729  0.435
## Retailer.countryUnited States   -0.015325  0.026044 -0.588
## Year                           0.032091  0.008830  3.634
##                                         Pr(>|t|)
## (Intercept)                      0.000352 ***
## prodCosts                         < 2e-16 ***
## Product.lineGolf Equipment        < 2e-16 ***
## Product.lineMountaineering Equipment 8.95e-10 ***
## Product.lineOutdoor Protection    < 2e-16 ***
## Product.linePersonal Accessories  4.72e-07 ***
## Retailer.countryAustria           0.068785 .
## Retailer.countryBelgium            0.029989 *
## Retailer.countryBrazil             0.095720 .
## Retailer.countryCanada             0.479103
## Retailer.countryChina              0.762088
## Retailer.countryDenmark            0.007646 **
## Retailer.countryFinland            0.328753
## Retailer.countryFrance             0.584973
## Retailer.countryGermany            0.559935
## Retailer.countryItaly               0.183607
## Retailer.countryJapan              0.697021
## Retailer.countryKorea              0.462324
## Retailer.countryMexico              0.222533
## Retailer.countryNetherlands        0.228001
## Retailer.countrySingapore           0.383691
## Retailer.countrySpain               0.144177
## Retailer.countrySweden              0.029880 *
## Retailer.countrySwitzerland         0.966563
## Retailer.countryUnited Kingdom      0.664189
## Retailer.countryUnited States       0.556966
## Year                            0.000362 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05506 on 183 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 8.118e+04 on 26 and 183 DF,  p-value: < 2.2e-16

# predict the outcome of the test data set
model.rev.prod.pred <- predict(model.rev.prod, subset(revenue.by_prod.test, Year==2006))
# measure against test data
model.rev.prod.lmse <- sqrt(sum((model.rev.prod.pred -
                                    subset(revenue.by_prod.test, Year==2006)$revenues))^2)
model.rev.prod.lmse

## [1] 1.435359

```

The investigation into the inclusion of products costs by product line, country and year into the model shows a significant improvement. However, since we're adding more variables to the model the  $R^2$  will naturally increase, although  $R^2 = 0.9999$  and a highly significant F-statistic indicate a model that fits the data better. The LMSE has dropped to 1.435359 with this model which indicates we are not in danger of overfitting as yet.

Interestingly, the country significance levels have dropped to the point where we consider removing them from the model, which we do in the final step.

#### Model 4: Adding Planned Revenue, Removing Country

```
# create a simple OLS model for log(revenue) ~ Product.cost
model.rev.prod <- lm(revenues ~ plannedRev +
                      prodCosts + Product.line +
                      Year,
                      data=revenue.by_prod.train )
summary(model.rev.prod)

##
## Call:
## lm(formula = revenues ~ plannedRev + prodCosts + Product.line +
##      Year, data = revenue.by_prod.train)
##
## Residuals:
##       Min     1Q     Median     3Q     Max
## -0.42812 -0.01490 -0.00276  0.02408  0.10564
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)                 -5.748e+01  1.658e+01 -3.467
## plannedRev                  -1.867e-09  3.797e-10 -4.916
## prodCosts                   1.036e+00  9.877e-04 1048.915
## Product.lineGolf Equipment   1.872e-01  1.222e-02 15.316
## Product.lineMountaineering Equipment 8.157e-02  1.424e-02  5.728
## Product.lineOutdoor Protection 5.338e-01  1.320e-02 40.432
## Product.linePersonal Accessories 6.575e-02  1.164e-02  5.651
## Year                         2.863e-02  8.274e-03  3.460
##
## Pr(>|t|)
## (Intercept)          0.000643 ***
## plannedRev           1.83e-06 ***
## prodCosts            < 2e-16 ***
## Product.lineGolf Equipment < 2e-16 ***
## Product.lineMountaineering Equipment 3.65e-08 ***
## Product.lineOutdoor Protection < 2e-16 ***
## Product.linePersonal Accessories 5.38e-08 ***
## Year                  0.000658 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05325 on 202 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 3.224e+05 on 7 and 202 DF,  p-value: < 2.2e-16

# predict the outcome of the test data set
model.rev.prod.pred <-
  predict(model.rev.prod,
         subset(revenue.by_prod.test, Year==2006))
# measure against test data
```

```
model.rev.prod.lmse <-
  sqrt(sum((model.rev.prod.pred -
            subset(revenue.by_prod.test, Year==2006)$revenues))^2)
model.rev.prod.lmse

## [1] 0.445751
```

This model removes the country variable but adds in the Planned.revenue variable. The model is improved with a LMSE of 0.445751 and highly significant F-statistic. The model is simplified by the removal of the country variable as well, helping to reduce variance while keeping bias as low as possible.