# **W271**-2 – Spring 2016 – **Lab 2**

**Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song**

March 7, 2016

## Contents

**Question 1: Broken Rulers** **2**

**Question 2: Investing** **3**

**Question 3: Turtles** **4**

**Question 4: CLM 1** **5**

Background . . . . . . . . 5

The Data . . . . . . . . 5

Question 4.1 . . . . . . . . 5

Question 4.2 . . . . . . . . 9

Question 4.3 . . . . . . . . 16

Question 4.4 . . . . . . . . 17

Question 4.5 . . . . . . . . 18

Question 4.6 . . . . . . . . 18

**Question 5: CLM 2** **20**

**Question 6: CLM 3** **27**

```
## Warning: package 'stargazer' was built under R version 3.1.3

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

# 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, $E(Y|X)$.**
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.**
3. **Write down an expression for the joint probability density function of X and Y, $f_{X,Y}(x, y)$.**
4. **Find the conditional probability density function of X given Y, $f_{X|Y}$.**
5. **Find the expectation of X, given that Y is 1/2, $E(X|Y = 1/2)$.**

---

# Question 2: Investing

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

**You plan to invest a total of one unit in all three companies. You will invest amount a in the first company, b in the second, and c in the third, where $\mathbf{a, b, c \in [0, 1]}$ and $\mathbf{a + b + c = 1}$. Find, the values of a, b, and c that minimize the variance of your total payoff.**

---

# Question 3: Turtles

**Next, suppose that the lifespan of a species of turtle follows a uniform distribution over $[\mathbf{0}, \theta]$. 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 $\mathbf{y_i}$.**

1. **Write down the likelihood function, $\mathbf{l}(\theta)$ in terms of $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}$.**
2. **Based on the previous result, what is the maximum-likelihood estimator for $\theta$?**
3. **Let $\hat{\theta_{\mathbf{ml}}}$ be the maximum likelihood estimator above. For the simple case that $\mathbf{n} \geq \mathbf{1}$, what is the expectation of $\hat{\theta_{\mathbf{ml}}}$, given $\theta$?**
4. **Is the maximum likelihood estimator biased?**
5. **For the more general case that $\mathbf{n} \geq \mathbf{1}$, what is the expectation of $\hat{\theta_{\mathbf{ml}}}$?**
6. **Is the maximum likelihood estimator consistent?**

---

# Question 4: CLM 1

## Background

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

## The Data

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

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

## Question 4.1

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

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

```
d<-read.csv("WageData2.csv")
summary(d)
```

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

```
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.000   Max.   :1.00   Max.   :1.000
##
##     IQscore         logWage
##  Min.   : 50.0   Min.   :4.844
##  1st Qu.: 93.0   1st Qu.:5.991
##  Median :103.0   Median :6.297
##  Mean   :102.3   Mean   :6.263
##  3rd Qu.:113.0   3rd Qu.:6.555
##  Max.   :144.0   Max.   :7.785
##  NA's   :316
```

```r
head(d)
```

```
##      X wage education experience age raceColor dad_education mom_education
## 1  191  951        12         10  28         0            NA            12
## 2 2059  288         8         11  25         1            NA             7
## 3 2072  509        12          6  24         0            12             9
## 4  945  647        18          5  29         0            12            12
## 5 1920  225        10         11  27         1             5             5
## 6 1927  454        10         11  27         1            NA             1
##   rural city z1 z2 IQscore  logWage
## 1     0    1  1  1     122 6.857514
## 2     1    0  0  1      NA 5.662960
## 3     1    1  0  0     127 6.232448
## 4     0    1  0  1     110 6.472346
## 5     1    0  0  1      NA 5.416100
## 6     1    0  0  1      NA 6.118097
```

```r
######### Function below to show multiple plots on a grid from  source http://www.cookbook-r.com/Graphs/

# Multiple plot function
#
# ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
# - cols:   Number of columns in layout
# - layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
# If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
# then plot 1 will go in the upper left, 2 will go in the upper right, and
# 3 will go all the way across the bottom.
#
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
```

```
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
###########

p1<-ggplot(d, aes(x=wage)) + geom_histogram(binwidth=100)
p2<-ggplot(d, aes(x=education)) + geom_histogram(binwidth=2)
p3<-ggplot(d, aes(x=experience)) + geom_histogram(binwidth=2)
p4<-ggplot(d, aes(x=age)) + geom_histogram(binwidth=1)
p5<-ggplot(d, aes(x=raceColor)) + geom_histogram(binwidth=.5)
p6<-ggplot(d, aes(x=dad_education)) + geom_histogram(binwidth=1)
p7<-ggplot(d, aes(x=mom_education)) + geom_histogram(binwidth=1)
p8<-ggplot(d, aes(x=rural)) + geom_histogram(binwidth=.5)
p9<-ggplot(d, aes(x=city)) + geom_histogram(binwidth=.5)

multiplot(p1, p2, p3, p4, p5, p6, p7, p8, p9, cols=3)
```

```
d$logWage<-log(d$wage)
d$experienceSquare<-d$experience^2
```

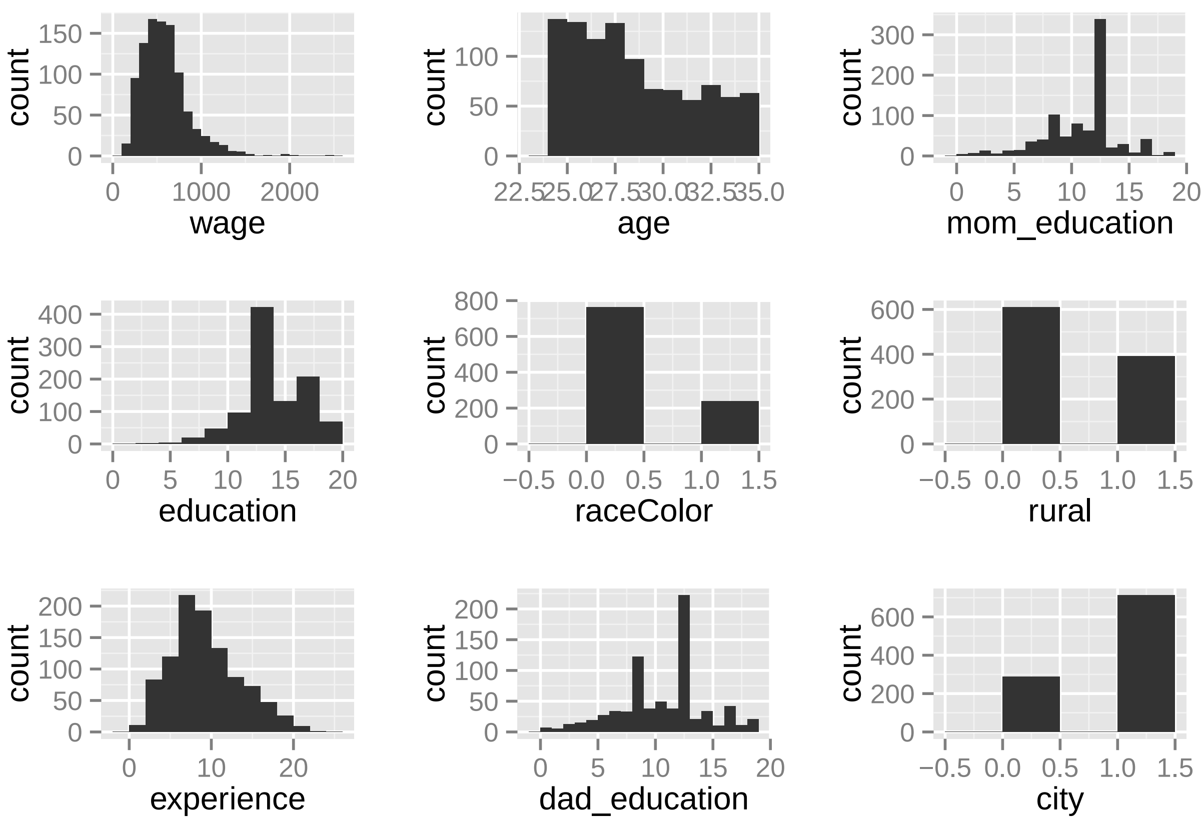
count
150
100
50
0
0
1000
2000
wage
count
100
50
0
22.5
25.0
27.5
30.0
32.5
35.0
age
count
300
200
100
0
0
5
10
15
20
mom_education
count
400
300
200
100
0
0
5
10
15
20
education
count
800
600
400
200
0
−0.5
0.0
0.5
1.0
1.5
raceColor
count
600
400
200
0
−0.5
0.0
0.5
1.0
1.5
rural
count
200
150
100
50
0
0
10
20
experience
count
200
150
100
50
0
0
5
10
15
20
dad_education
count
600
400
200
0
−0.5
0.0
0.5
1.0
1.5
city

Figure 1:

### 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.**

```
p2.1<-ggplot(d, aes(wage, education)) +
  geom_point() +
  labs(x = "Wage",
       y = "Education") +
  geom_smooth(method = "lm")

p2.2<-ggplot(d, aes(log(wage), education)) +
  geom_point() +
  labs(x = "Log(Wage)",
       y = "Education") +
  geom_smooth(method = "lm")

p2.3<-ggplot(d, aes(wage, experience)) +
  geom_point() +
  labs(x = "Wage",
       y = "Experience") +
  geom_smooth(method = "lm")

p2.4<-ggplot(d, aes(log(wage), experience)) +
  geom_point() +
  labs(x = "Log(Wage)",
       y = "Experience") +
  geom_smooth(method = "lm")

p2.5<-ggplot(d, aes(wage, age)) +
  geom_point() +
  labs(x = "Wage",
       y = "Age") +
  geom_smooth(method = "lm")

p2.6<-ggplot(d, aes(log(wage), age)) +
  geom_point() +
  labs(x = "Log(Wage)",
       y = "Age") +
  geom_smooth(method = "lm")

p2.7<-ggplot(d, aes(wage, raceColor)) +
  geom_point() +
  labs(x = "Wage",
       y = "Race") +
  geom_smooth(method = "lm")

p2.8<-ggplot(d, aes(log(wage), raceColor)) +
  geom_point() +
  labs(x = "Log(Wage)",
       y = "Race") +
  geom_smooth(method = "lm")

p2.9<-ggplot(d, aes(wage, dad_education)) +
```

```r
  geom_point(na.rm = T) +
  labs(x = "Wage",
       y = "Father's Education") +
  geom_smooth(method = "lm", na.rm = T)

p2.10<-ggplot(d, aes(log(wage), dad_education)) +
  geom_point(na.rm = T) +
  labs(x = "Log(Wage)",
       y = "Father's Education") +
  geom_smooth(method = "lm", na.rm = T)

p2.11<-ggplot(d, aes(wage, mom_education)) +
  geom_point(na.rm = T) +
  labs(x = "Wage",
       y = "Mother's Education") +
  geom_smooth(method = "lm", na.rm = T)

p2.12<-ggplot(d, aes(log(wage), mom_education)) +
  geom_point(na.rm = T) +
  labs(x = "Log(Wage)",
       y = "Mother's Education") +
  geom_smooth(method = "lm", na.rm = T)

p2.13<-ggplot(d, aes(wage, rural)) +
  geom_point(na.rm = T) +
  labs(x = "Wage",
       y = "Location - Rural") +
  geom_smooth(method = "lm", na.rm = T)

p2.14<-ggplot(d, aes(log(wage), rural)) +
  geom_point(na.rm = T) +
  labs(x = "Log(Wage)",
       y = "Location - Rural") +
  geom_smooth(method = "lm", na.rm = T)

p2.15<-ggplot(d, aes(wage, city)) +
  geom_point(na.rm = T) +
  labs(x = "Wage",
       y = "Location - City") +
  geom_smooth(method = "lm", na.rm = T)

p2.16<-ggplot(d, aes(log(wage), city)) +
  geom_point(na.rm = T) +
  labs(x = "Log(Wage)",
       y = "Location - City") +
  geom_smooth(method = "lm", na.rm = T)

p2.17<-ggplot(d, aes(wage, IQscore)) +
  geom_point(na.rm = T) +
  labs(x = "Wage",
       y = "IQ") +
  geom_smooth(method = "lm", na.rm = T)
```

```
p2.18<-ggplot(d, aes(log(wage), IQscore)) +
  geom_point(na.rm = T) +
  labs(x = "Log(Wage)",
       y = "IQ") +
  geom_smooth(method = "lm", na.rm = T)


multiplot(p2.1, p2.2, p2.3, p2.4, cols=2)
```

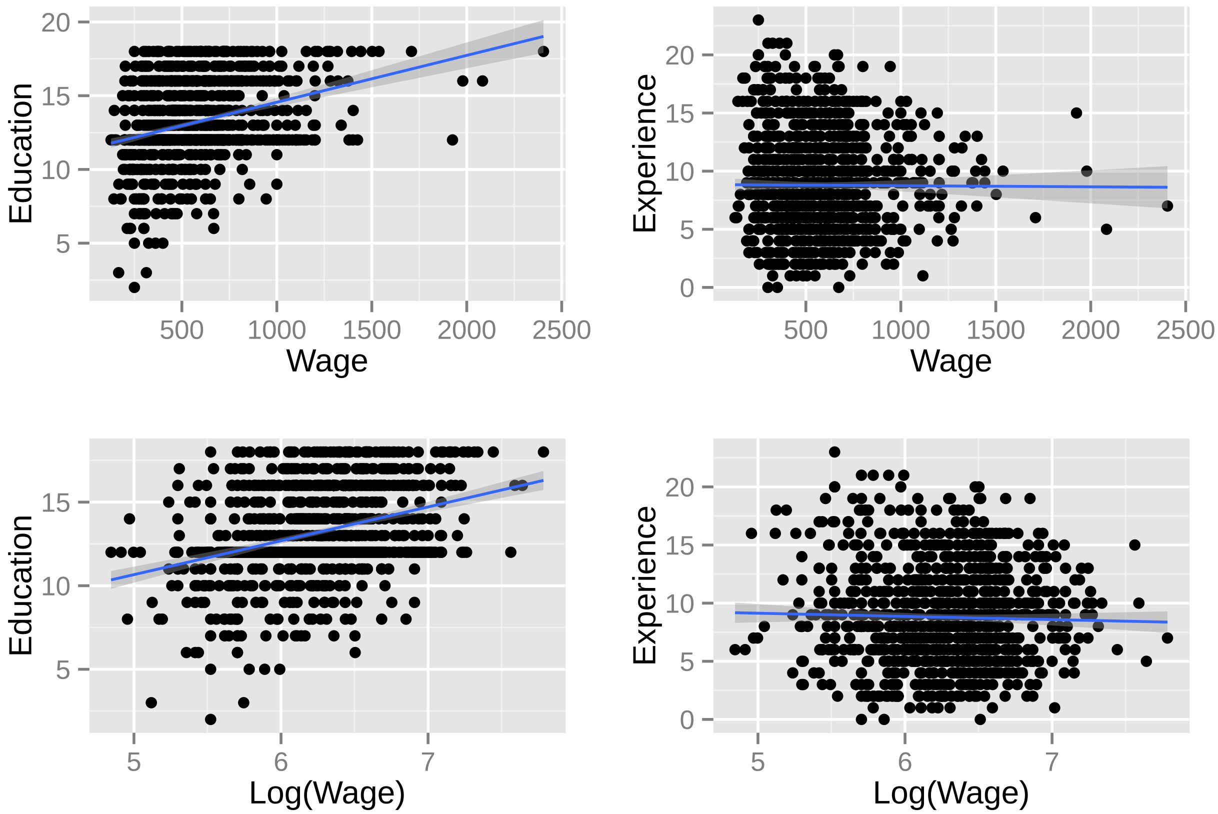


Figure 2:

```
multiplot(p2.5, p2.6, p2.7, p2.8, cols=2)
```

```
multiplot(p2.9,p2.10, p2.11, p2.12, cols=2)
```

```
multiplot(p2.13, p2.14, p2.15, p2.16, cols=2)
```

```
multiplot(p2.17, p2.18, cols=2)
```

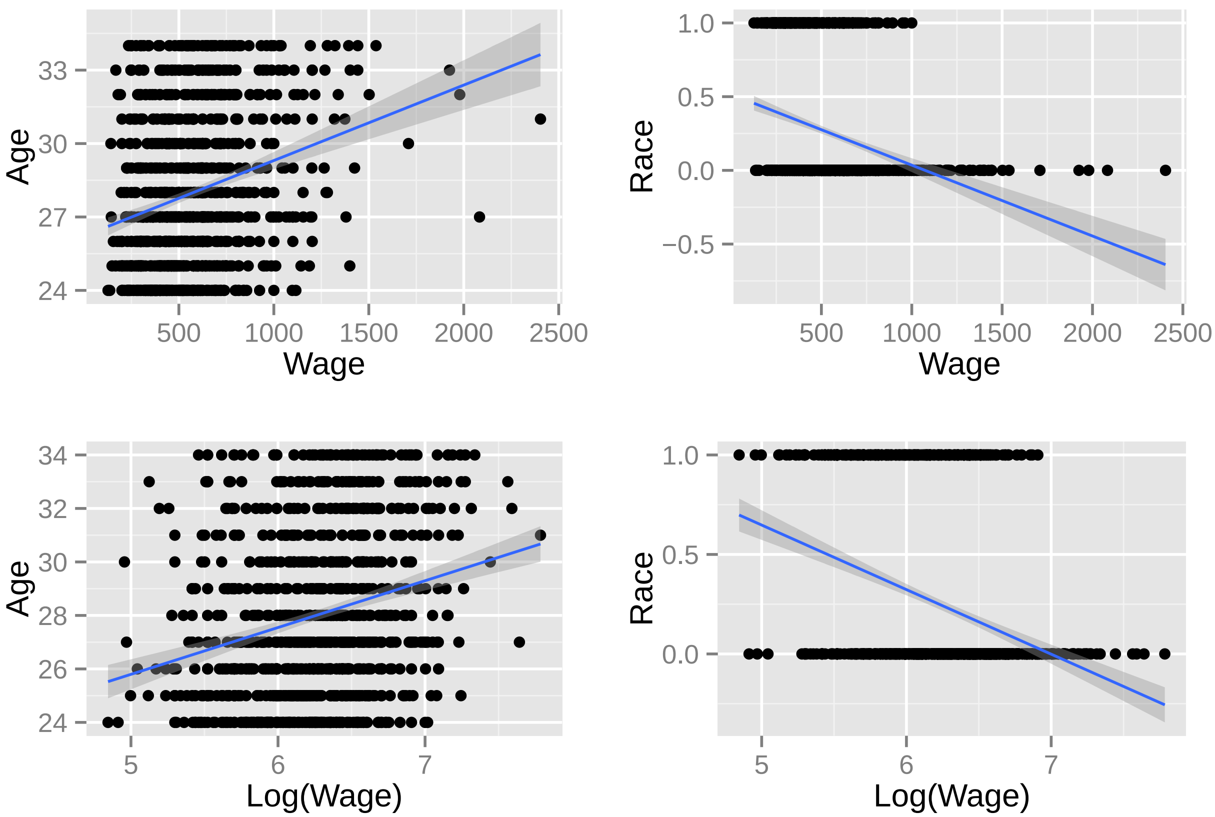


Figure 3:

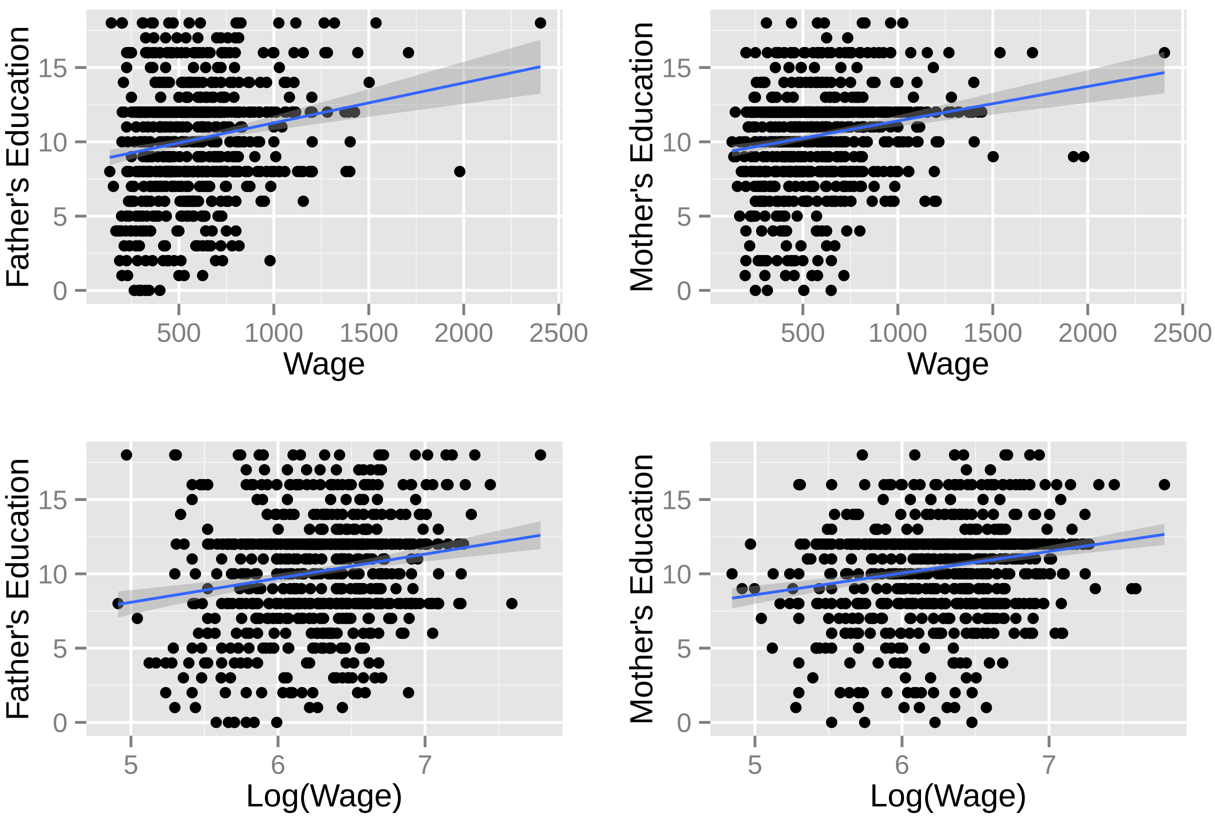
Father's Education
Wage
Mother's Education
Wage
Father's Education
Log(Wage)
Mother's Education
Log(Wage)

Figure 4:

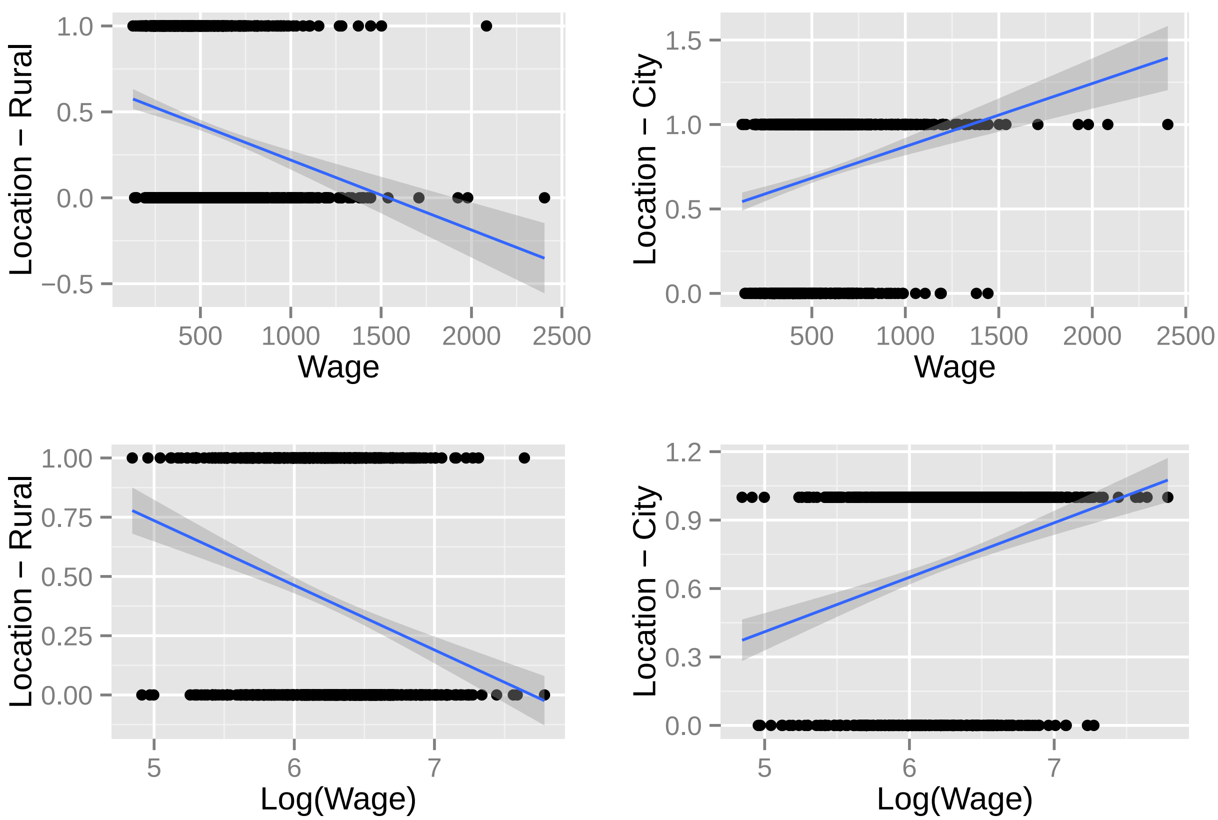
Location – Rural
Wage
Location – City
Wage
Location – Rural
Log(Wage)
Location – City
Log(Wage)

Figure 5:

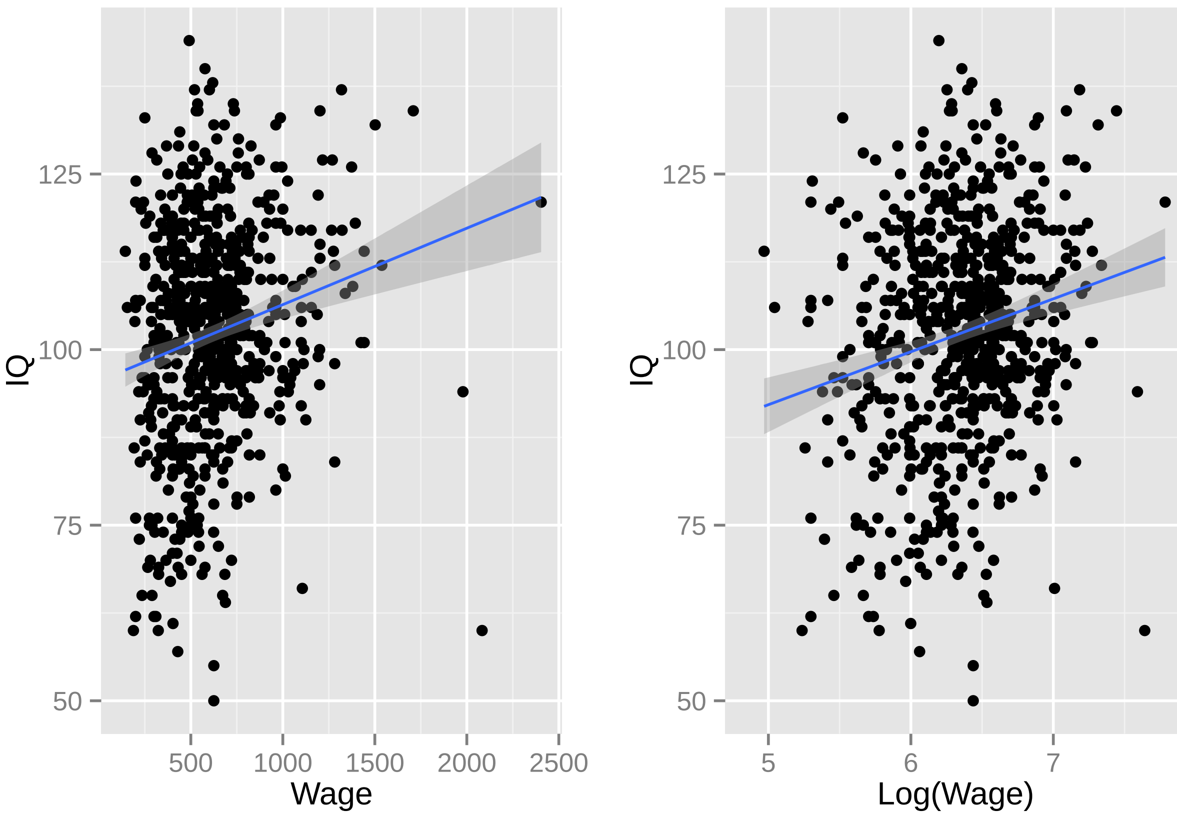
IQ
125
100
75
50
500
1000
1500
2000
2500
Wage
IQ
125
100
75
50
5
6
7
Log(Wage)

Figure 6:

## Question 4.3

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

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

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

```
stargazer(model4.3, type="latex", title="Question 4.3-1")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Mar 01, 2016 - 1:33:47 PM

Table 1: Question 4.3-1

| | *Dependent variable:* |
|---|---|
| | logWage |
| education | 0.080*** |
| | (0.006) |
| experience | 0.035*** |
| | (0.004) |
| age | |
| raceColor | −0.261*** |
| | (0.030) |
| Constant | 4.962*** |
| | (0.113) |
| Observations | 1,000 |
| $R^2$ | 0.236 |
| Adjusted $R^2$ | 0.234 |
| Residual Std. Error | 0.392 (df = 996) |
| F Statistic | 102.582*** (df = 3; 996) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

2. **Explain why the degrees of freedom takes on the specific value you observe in the regression output.** The degrees of freedom on the residual errors is $1000 - 4 = 996$ where the 1000 is the number of observations and 4 is subtracted for th enumber of covariates included in the model. For the F-statistic the degrees of freedom is $4000 - 4 = 3996$ where the 4000 accounts for the 1000 observations for each covariate and the 4 is again for the four covariates in the model.

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).**

There is no output for the age variable and

4. **Interpret the coefficient estimate associated with education.**

The coefficient on education is $0.080 \pm 0.012$ which indicates that for 1 additional year of education wage would increase by an estimated 8 percent.

5. **Interpret the coefficient estimate associated with experience.**

The coefficient on experinece is $0.035 \pm 0.008$ which indicates that for 1 additional year of experience wage would increase by an estimated 3.5 percent.

## Question 4.4

**Regress *log(wage)* on education, experience, experienceSquare, and race-Color.**

```
model4.4<-lm(logWage~education + experience + experienceSquare + raceColor, data=d)
stargazer(model4.3, model4.4, type="latex", title="Question 4.4")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Mar 01, 2016 - 1:33:47 PM

Table 2: Question 4.4

| | *Dependent variable:* | |
|---|---|---|
| | logWage | |
| | (1) | (2) |
| education | 0.080*** | 0.079*** |
| | (0.006) | (0.006) |
| experience | 0.035*** | 0.092*** |
| | (0.004) | (0.012) |
| age | | |
| experienceSquare | | −0.003*** |
| | | (0.001) |
| raceColor | −0.261*** | −0.263*** |
| | (0.030) | (0.030) |
| Constant | 4.962*** | 4.736*** |
| | (0.113) | (0.120) |
| Observations | 1,000 | 1,000 |
| $R^2$ | 0.236 | 0.257 |
| Adjusted $R^2$ | 0.234 | 0.254 |
| Residual Std. Error | 0.392 (df = 996) | 0.387 (df = 995) |
| F Statistic | 102.582*** (df = 3; 996) | 85.978*** (df = 4; 995) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

1. **Plot a graph of the estimated effect of experience on wage.**
2. **What is the estimated effect of experience on wage when experience is 10 years?**

## 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=d)
stargazer(model4.3, model4.4, model4.5, type="latex", title="Question 4.5")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Mar 01, 2016 - 1:33:47 PM

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

The number of observations is only 723 out of the total 1000 so missing values are an issue.

```
#identify rows that have an NA value
row.has.na <- apply(d, 1, function(x){any(is.na(x))})

#make a dataframe of only those rows with missing values to look at
d_missing<-d[row.has.na,]
```

2. **Do you just want to "throw away" these observations?**
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?**
4. **How about regress the variable(s) with missing values on education, experience, and raceColor, and use this regression(s) to predict (i.e., "impute") the missing values and then rerun the original regression using all of the observations?**
5. **Compare the results of all of these regressions. Which one, if at all, would you prefer?**

## 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)?**
2. **Suppose $z_1$ is an indicator representing whether or not an individual lives in an area in which there was a recent policy change to promote the importance of education. Could $z_1$ be correlated with other unobservables captured in the error term?**
3. **Using the same specification as that in Question 4.5, estimate the equation by 2SLS, using both $z_1$ and $z_2$ as instrument variables. Interpret the results. How does the coefficient estimate on education change?**

Table 3: Question 4.5

| | *Dependent variable:* | | |
|---|---|---|---|
| | logWage | | |
| | (1) | (2) | (3) |
| education | 0.080*** | 0.079*** | 0.068*** |
| | (0.006) | (0.006) | (0.008) |
| experience | 0.035*** | 0.092*** | 0.097*** |
| | (0.004) | (0.012) | (0.013) |
| age | | | |
| experienceSquare | | −0.003*** | −0.003*** |
| | | (0.001) | (0.001) |
| raceColor | −0.261*** | −0.263*** | −0.213*** |
| | (0.030) | (0.030) | (0.043) |
| dad_education | | | −0.001 |
| | | | (0.005) |
| mom_education | | | 0.011* |
| | | | (0.006) |
| rural | | | −0.092*** |
| | | | (0.031) |
| city | | | 0.178*** |
| | | | (0.032) |
| Constant | 4.962*** | 4.736*** | 4.642*** |
| | (0.113) | (0.120) | (0.141) |
| Observations | 1,000 | 1,000 | 723 |
| $R^2$ | 0.236 | 0.257 | 0.275 |
| Adjusted $R^2$ | 0.234 | 0.254 | 0.267 |
| Residual Std. Error | 0.392 (df = 996) | 0.387 (df = 995) | 0.379 (df = 714) |
| F Statistic | 102.582*** (df = 3; 996) | 85.978*** (df = 4; 995) | 33.793*** (df = 8; 714) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

# Question 5: CLM 2

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

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

To start off with we need to examine the vairables and omit outliers. In this case there are some NA values in the absolute_wealth variable and there are a lot of data points very close to zero, which screw the data. This first model omits those entries. Additionally we transform the absolute_wealth into a log(absolute_wealth) to rescale the values for regression.

```
data<-read.csv('wealthy_candidates.csv')

summary(data)
```

```
##        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
```

```
head(data)
```

```
##   X   region        urb       lit voteshare absolute_wealth
## 1 1 Region 2 0.14909884 0.4283742 0.4168488      5110593.00
## 2 2 Region 2 0.14909884 0.4283742 0.1137623        99999.97
## 3 3 Region 2 0.09182214 0.4579071 0.2983904        55340.00
## 4 4 Region 2 0.10168768 0.3063438 0.4835877       206999.94
## 5 5 Region 2 0.06139975 0.2731756 0.3106902      1307408.00
## 6 6 Region 2 0.41726938 0.5199646 0.4023529      5864785.50
```

```
#look at the variables
hist(data$voteshare)
```

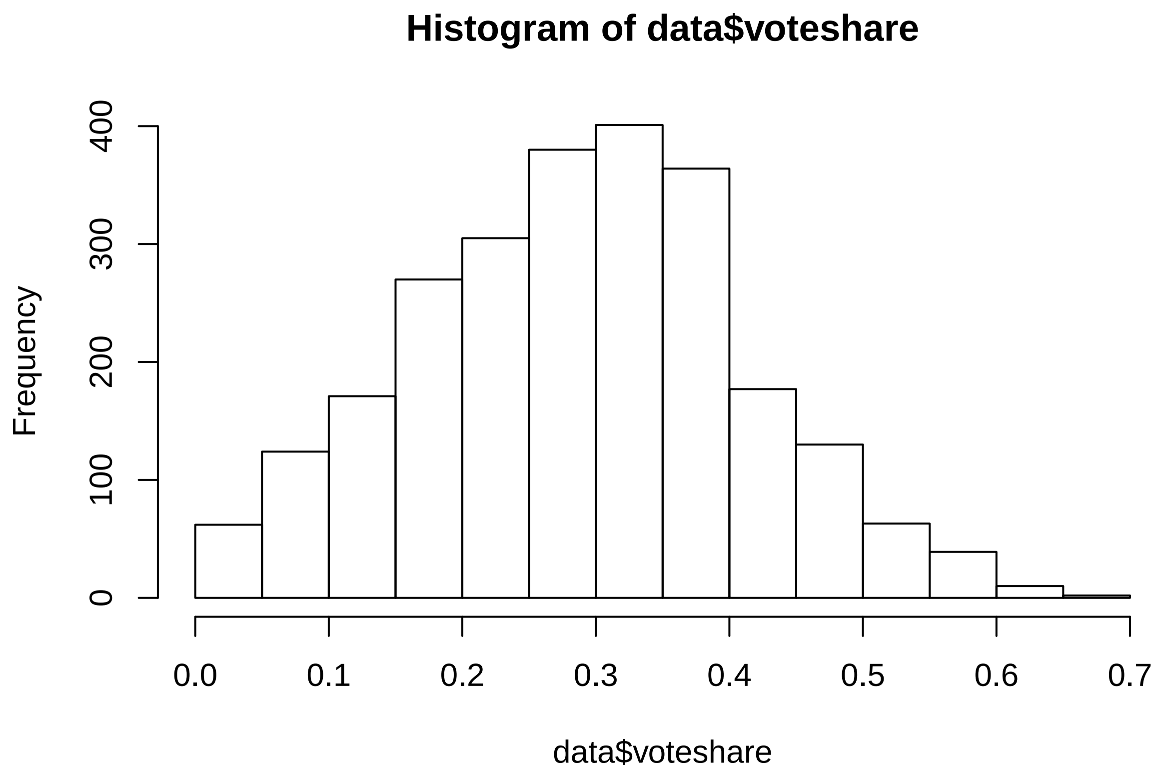


Figure 7:

```
hist(data$absolute_wealth)
```

```
#subset the data and only keep the complete cases
data2<-data[, c("voteshare", "absolute_wealth", "region")]
data2<-data2[complete.cases(data),]

data2$logwealth<-log(data2$absolute_wealth)
min(data2$logwealth)
```

```
## [1] 0.6931472
```

```
hist(data2$logwealth, breaks=20)
```

```
#eliminate the large column on near 0 absoluate wealth
data2<-subset(data2, logwealth>1, )

#plot the data to get a sense if there might be a linear relationship
```

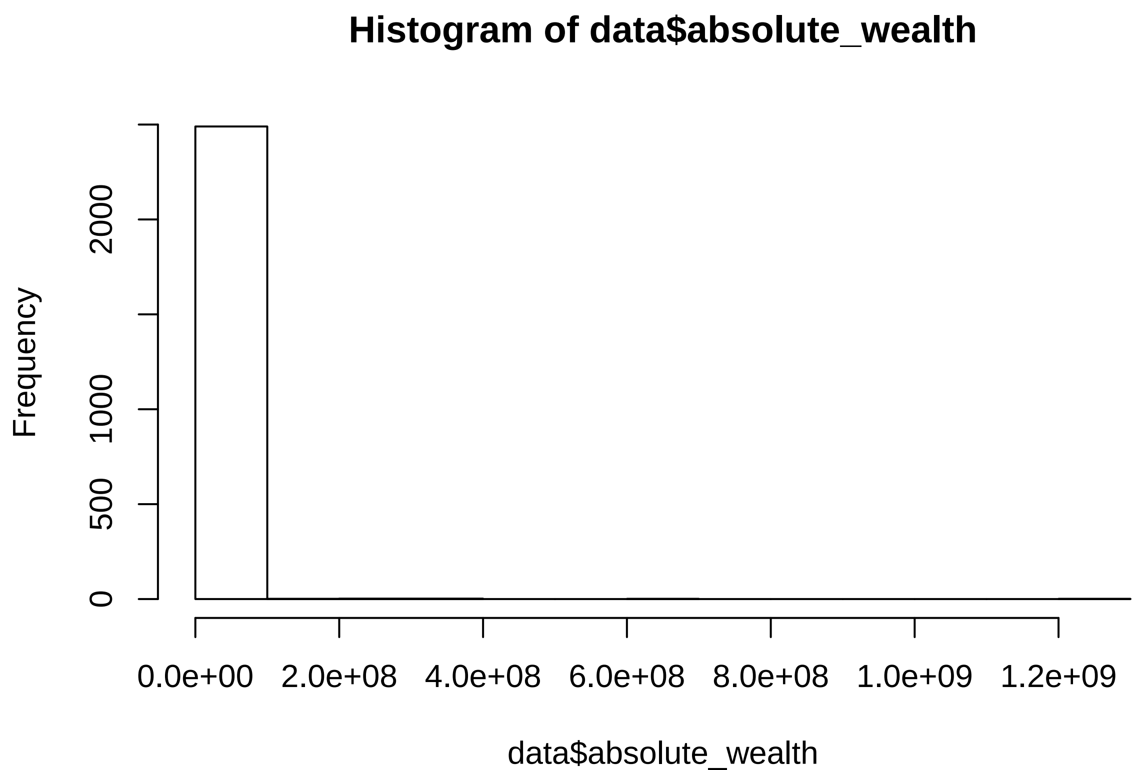
Histogram of data$absolute_wealth
Frequency
0
500
1000
2000
0.0e+00
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.0e+09
1.2e+09
data$absolute_wealth

Figure 8:

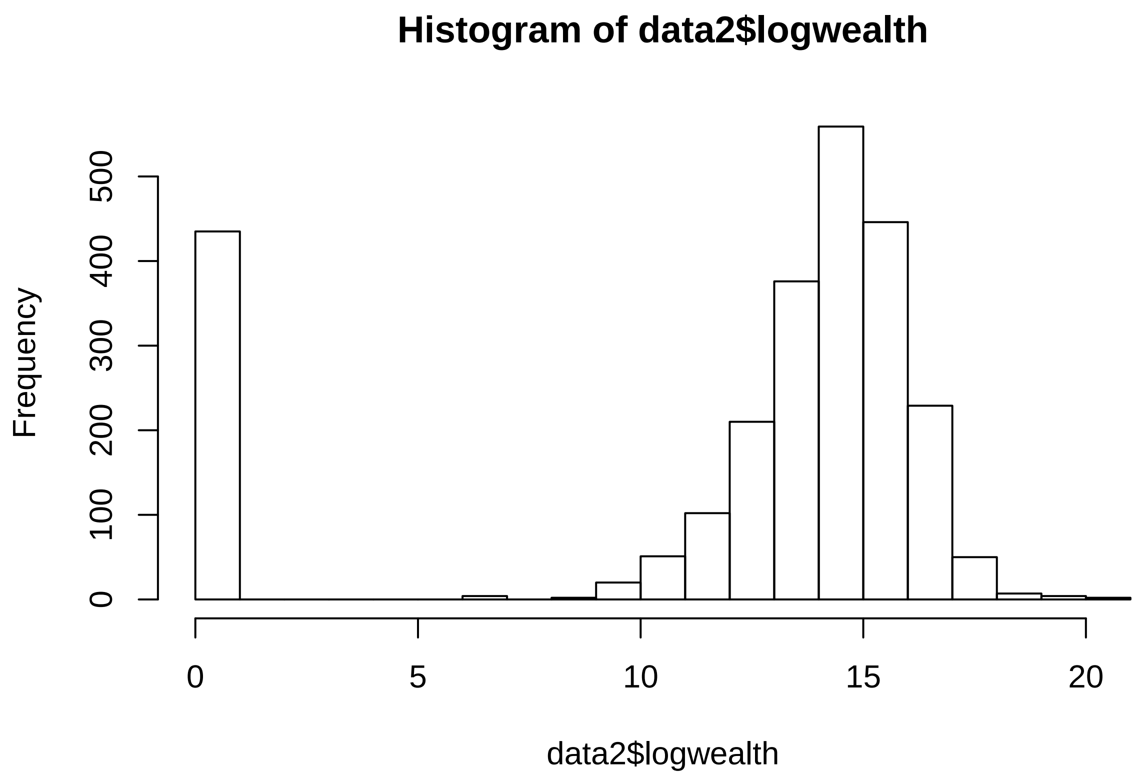
Histogram of data2$logwealth
Frequency
0
100
200
300
400
500
0
5
10
15
20
data2$logwealth

Figure 9:

```
scatter1<-ggplot(data2, aes(voteshare, logwealth))
scatter1+geom_point()
```

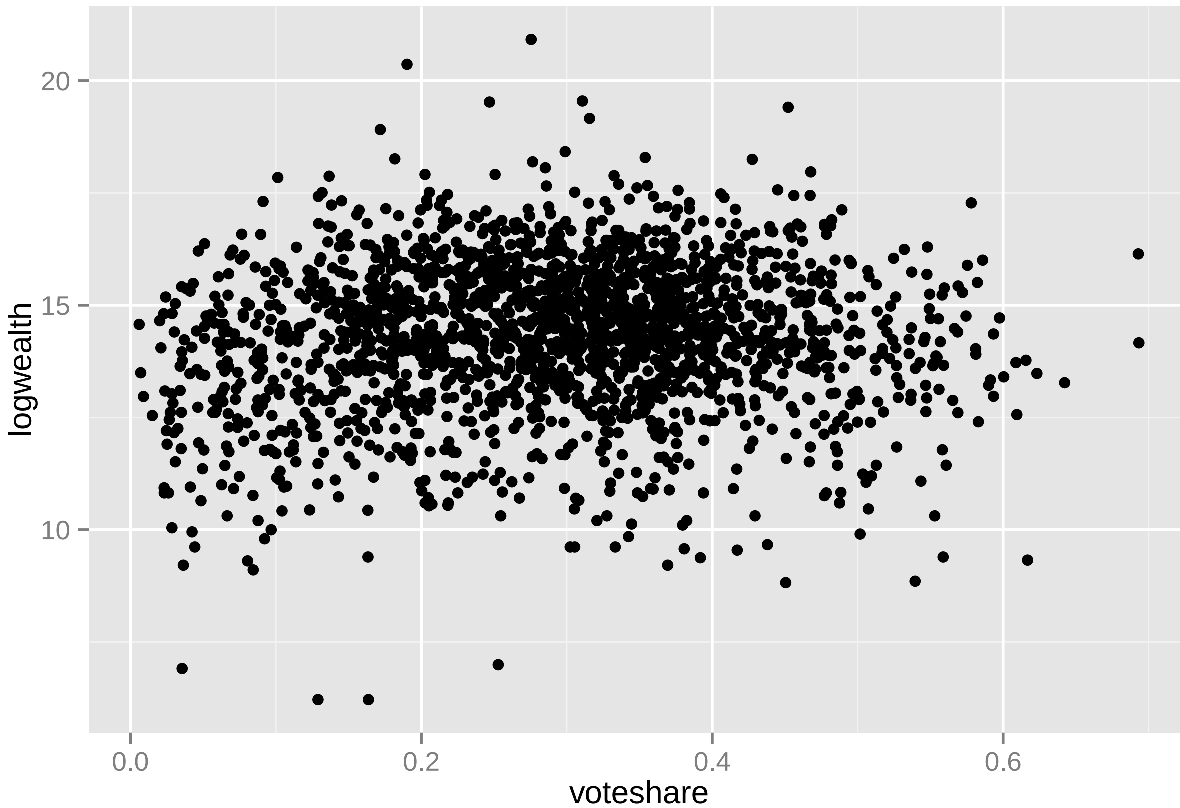


Figure 10:

```
#run the basic model
model1<-lm(voteshare~logwealth, data=data2)
stargazer(model1, type="latex", title="Question 5.1")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Mar 01, 2016 - 1:33:48 PM

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?**

```
#run the model with the squared term added in
model2<-lm(voteshare~logwealth+ logwealth*logwealth, data=data2)
stargazer(model1, model2, type="latex", title="Question 5.2")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Mar 01, 2016 - 1:33:48 PM

Table 4: Question 5.1

| | *Dependent variable:* |
|---|---|
| | voteshare |
| logwealth | 0.005*** |
| | (0.002) |
| Constant | 0.216*** |
| | (0.024) |
| Observations | 2,062 |
| $R^2$ | 0.005 |
| Adjusted $R^2$ | 0.004 |
| Residual Std. Error | 0.125 (df = 2060) |
| F Statistic | 9.513*** (df = 1; 2060) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Table 5: Question 5.2

| | *Dependent variable:* | |
|---|---|---|
| | voteshare | |
| | (1) | (2) |
| logwealth | 0.005*** | 0.005*** |
| | (0.002) | (0.002) |
| Constant | 0.216*** | 0.216*** |
| | (0.024) | (0.024) |
| Observations | 2,062 | 2,062 |
| $R^2$ | 0.005 | 0.005 |
| Adjusted $R^2$ | 0.004 | 0.004 |
| Residual Std. Error (df = 2060) | 0.125 | 0.125 |
| F Statistic (df = 1; 2060) | 9.513*** | 9.513*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

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.**

```
model3<-lm(voteshare~logwealth+factor(region), data=data2)

stargazer(model1, model2, model3, type="latex", omit="region",
          add.lines = list(c("Region Fixed effects", "No", "No", "Yes")), title="Question 5.3")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Mar 01, 2016 - 1:33:48 PM

Table 6: Question 5.3

| | *Dependent variable:* | | |
|---|---|---|---|
| | voteshare | | |
| | (1) | (2) | (3) |
| logwealth | $0.005^{***}$ | $0.005^{***}$ | $0.012^{***}$ |
| | (0.002) | (0.002) | (0.002) |
| Constant | $0.216^{***}$ | $0.216^{***}$ | $0.088^{***}$ |
| | (0.024) | (0.024) | (0.028) |
| Region Fixed effects | No | No | Yes |
| Observations | 2,062 | 2,062 | 2,062 |
| $R^2$ | 0.005 | 0.005 | 0.039 |
| Adjusted $R^2$ | 0.004 | 0.004 | 0.038 |
| Residual Std. Error | 0.125 (df = 2060) | 0.125 (df = 2060) | 0.123 (df = 2058) |
| F Statistic | $9.513^{***}$ (df = 1; 2060) | $9.513^{***}$ (df = 1; 2060) | $28.109^{***}$ (df = 3; 2058) |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

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: CLM 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.**

---