W271-2 – Spring 2016 – **Lab 2**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

March 7, 2016

Contents

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

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

We start by conducting univariate analysis on the dataset to look for any issues in the data that may violate the assumptions for regression models and to intendify and make decisions about missing values or potential outliers. We also can create new transform variables that rescale the data such as creating the logwage varaible.

```
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
```

```
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
```

```
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
```

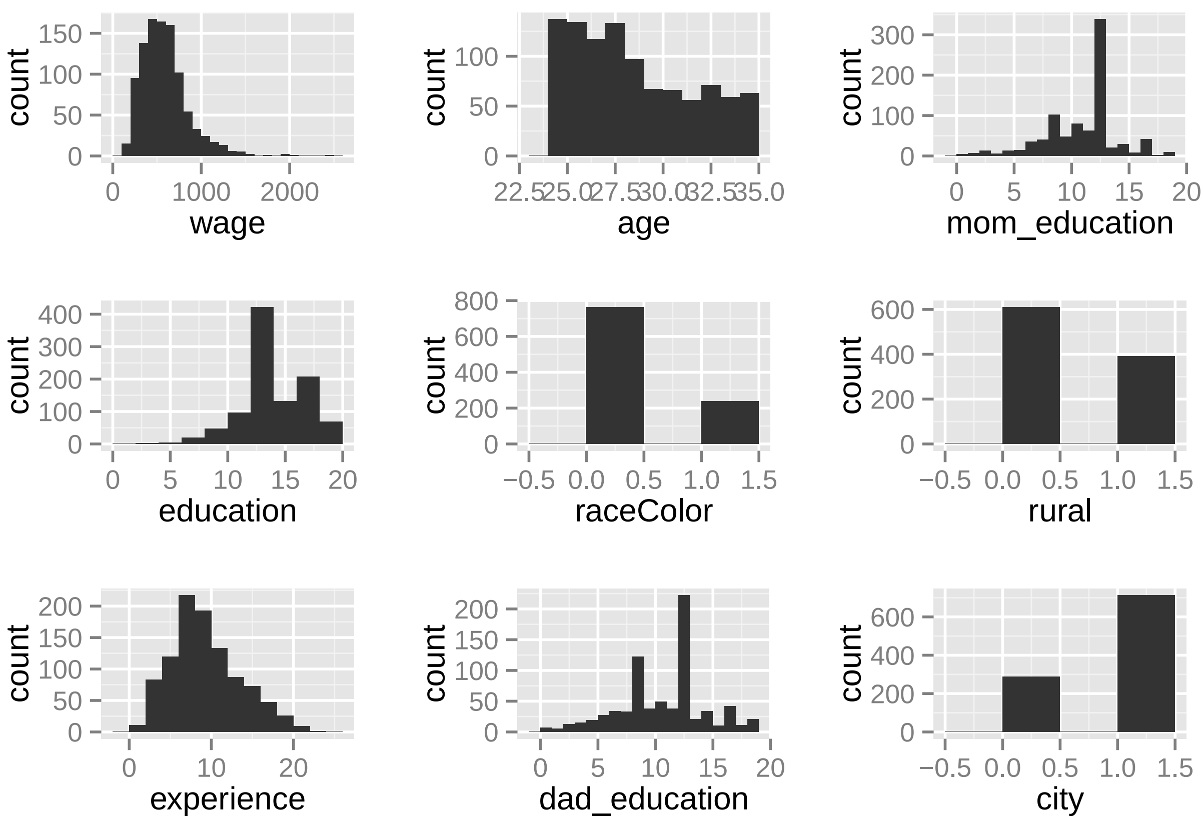

count
wage
count
age
count
mom_education
count
education
count
raceColor
count
rural
count
experience
count
dad_education
count
city

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.

We also conduct bivariate analysis to be able to identify potential patterns to inform our model and any additional outliers.

```
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)) +
  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)
```

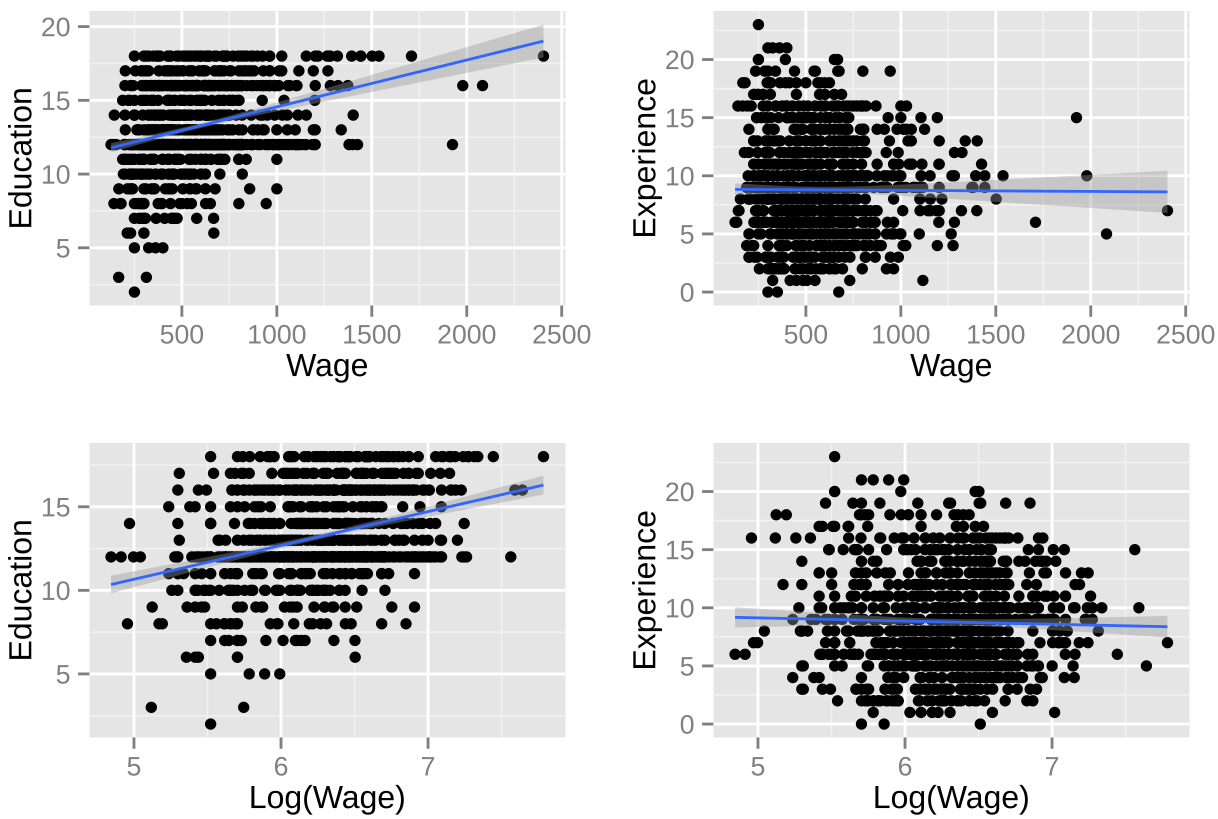


Figure 2:

```
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)
```

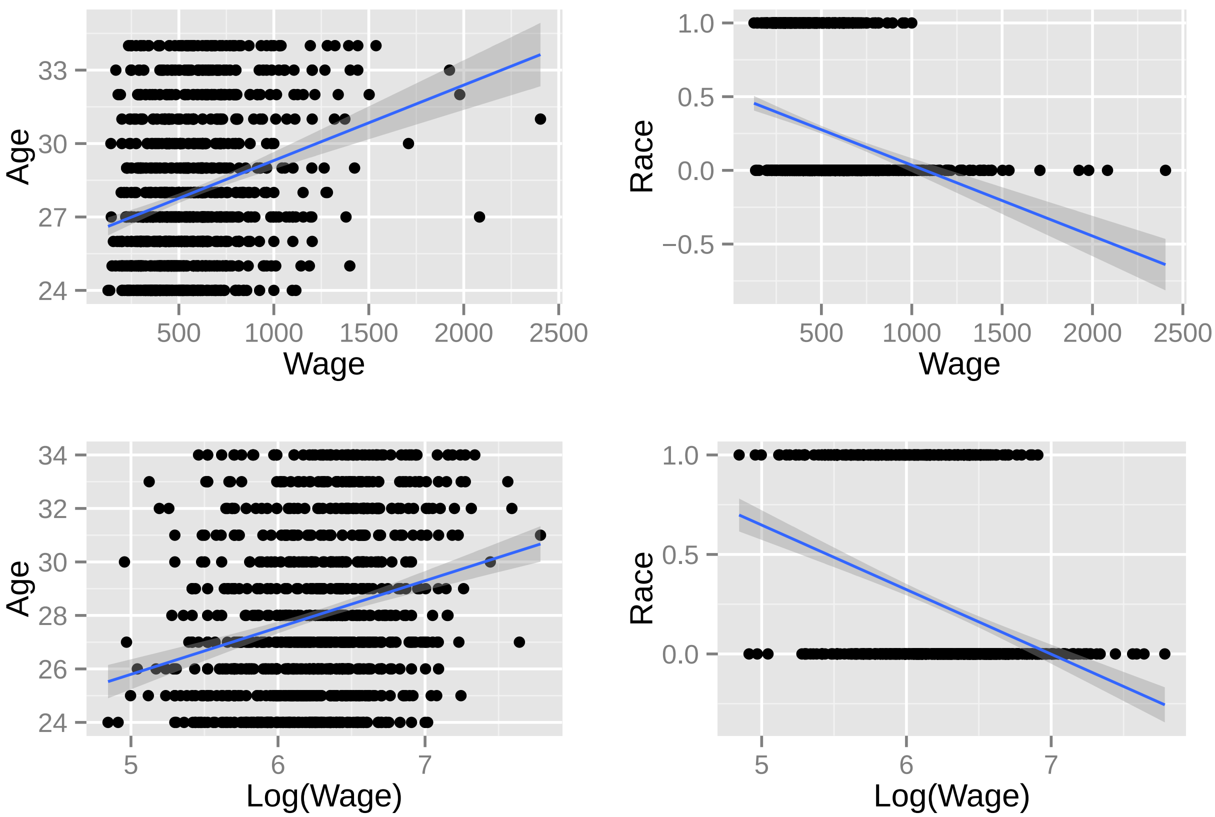


Figure 3:

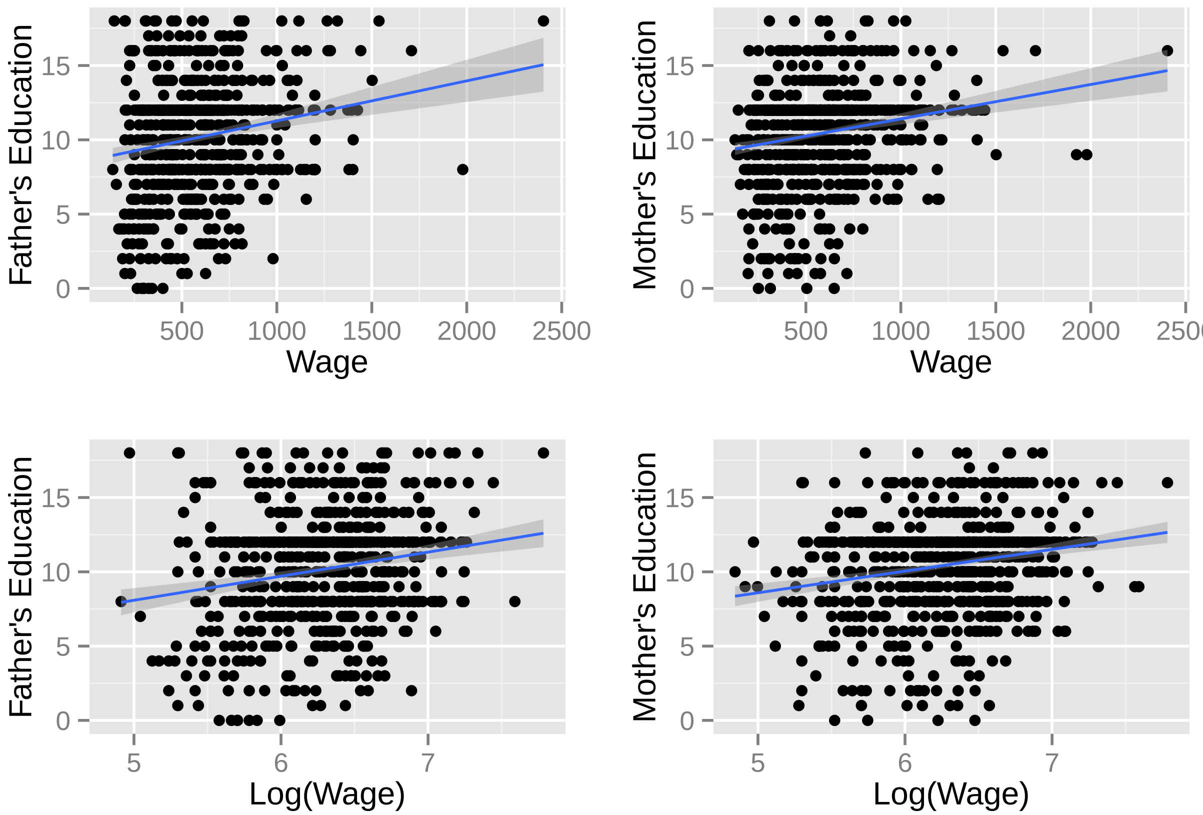
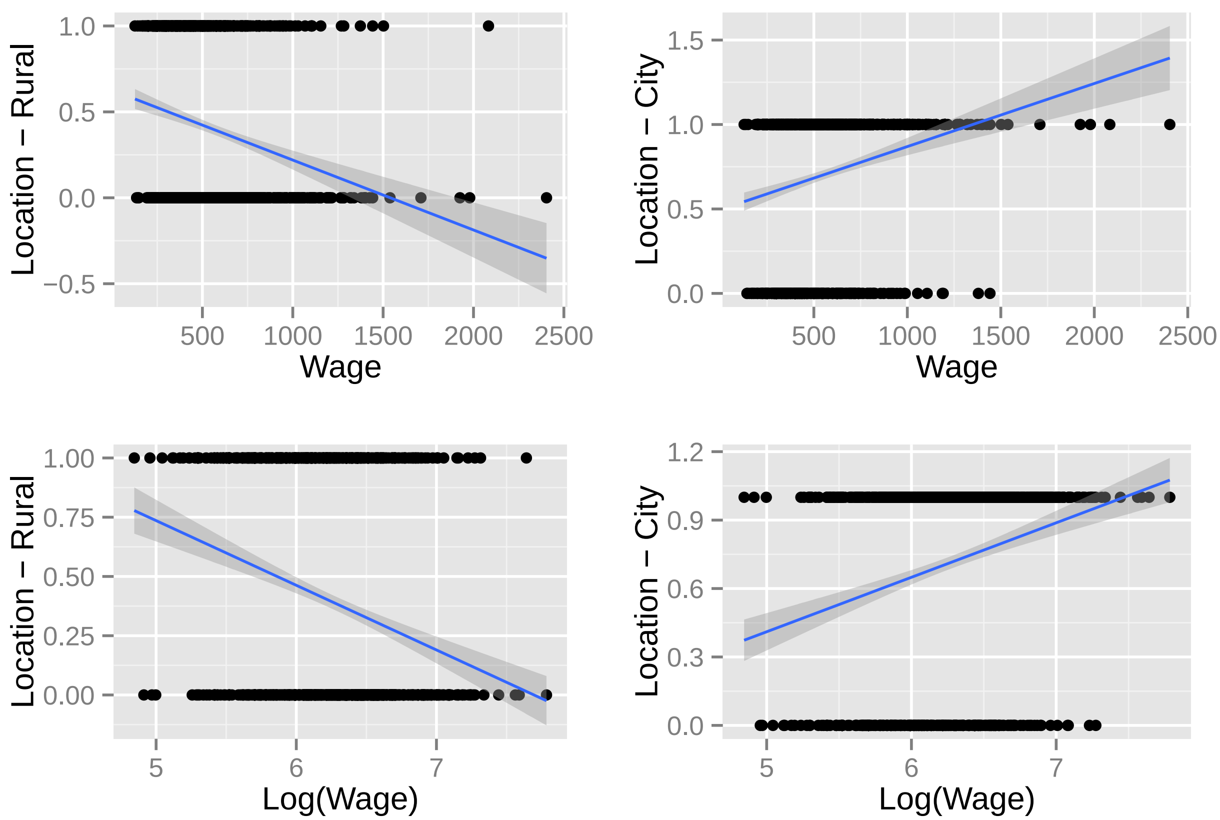

Father's Education
Wage
Mother's Education
Wage
Father's Education
Log(Wage)
Mother's Education
Log(Wage)

Figure 4:



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

Figure 5:

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

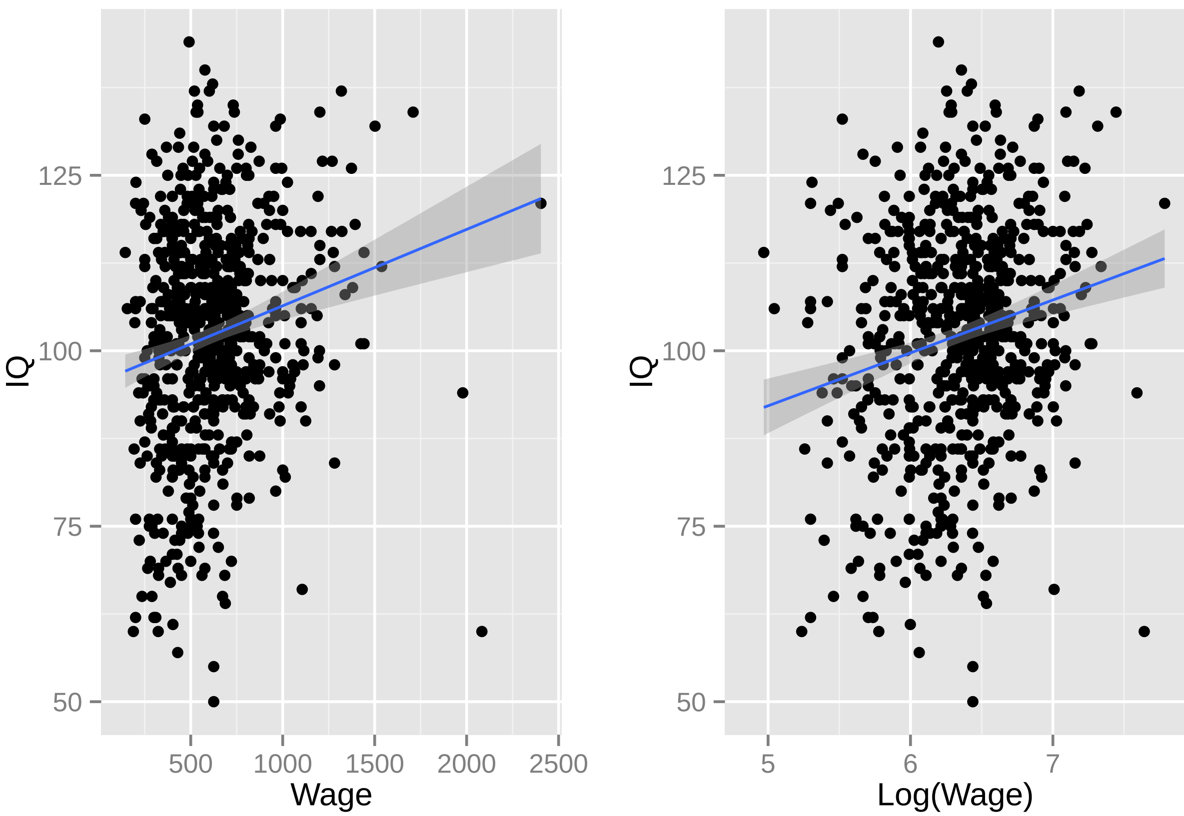


Figure 6:

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: Sun, Mar 06, 2016 - 5:05:54 PM

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

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

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.

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

The coefficient on education is 0.080 ± 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 ± 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: Sun, Mar 06, 2016 - 5:05:54 PM

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

```
#Note: I am not sure this is what is being asked for, not sure if I am interpreting the question correc


#pull out the coefficients from the model
coefs<-coef(model4.4)

#set x values to be the experence data
x<-d$experience

#set y to be just the effect from experience
#so we pull the ceofficient for expereience and for experienceSquared
y<-coefs[3]*x+coefs[4]*x^2

#put the data in a new dataframe to use for plotting
dat<-data.frame(x,y)

#plot the estimated effect of expereience (with the squared term) on the log(wage)
ggplot(dat, aes(x,y))+
  geom_smooth(na.rm=T) +
  labs(x="experience", y="log(wage)")
```

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

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, b
```

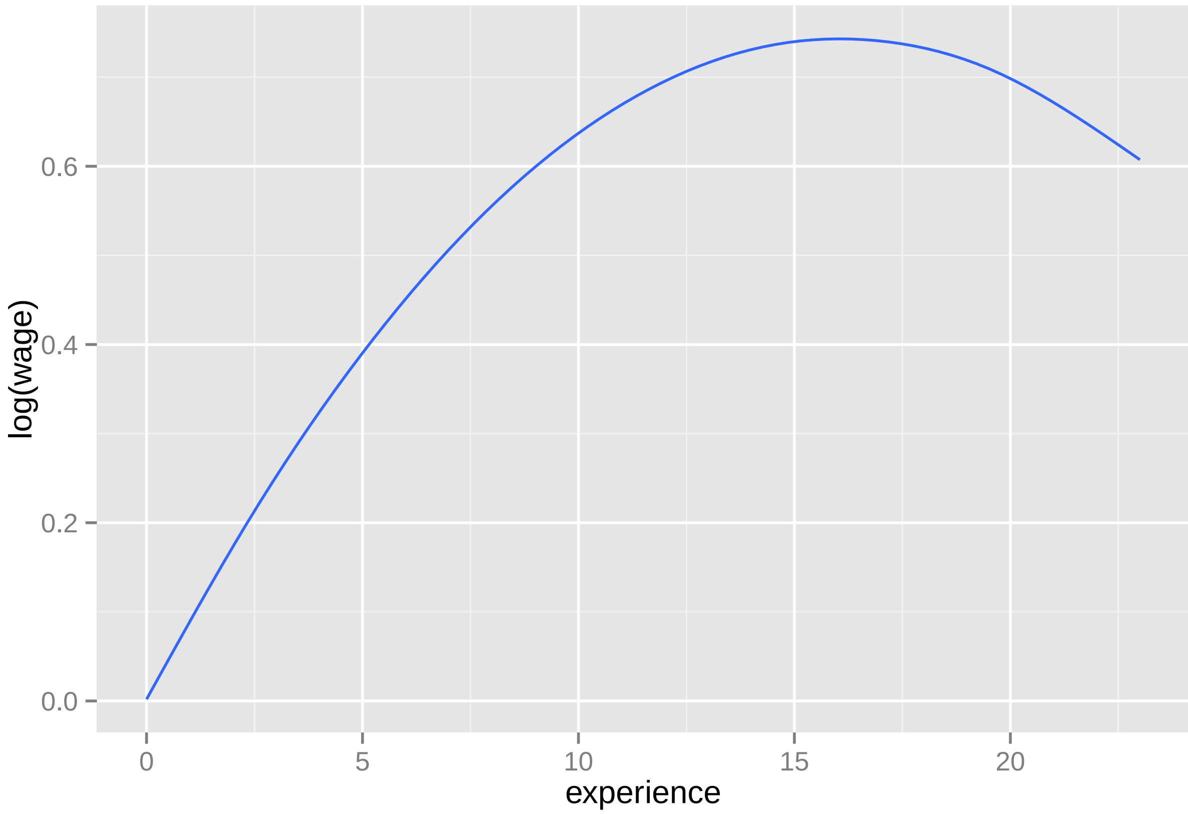


Figure 7:

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

When expereience equals 10 the estimated effect on wage is about 64 percent which can be read from the plot or derived from the coefficients $estimatedeffect = 0.0925 \cdot 10 + -0.003 \cdot 10^2 = 0.637$.

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: Sun, Mar 06, 2016 - 5:05:56 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.**

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

The number of observations is only 723 out of the total 1000 so there are missing values that may be causing issues. From the code below we see that the missing values are in the dad_education and mom_education variables. Those entries are seperated out to explore any patterns. The plots that follow explore the education, experience, and raceColor for those that are not complete cases (ie have missing data). No discernible pattern is present.

```
#check which variables have the missing data
sum(is.na(d$education))
```

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

```
sum(is.na(d$experience))
```

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

```
sum(is.na(d$experienceSquare))
```

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

```
sum(is.na(d$raceColor))
```

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

```
sum(is.na(d$dad_education))
```

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

```
sum(is.na(d$mom_education))
```

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

```
sum(is.na(d$rural))
```

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

```
sum(is.na(d$city))
```

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

```
#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,]
```

```r
#create plots to try and identify any patterns
p4.1<-ggplot(d_missing, aes(wage, education)) +
  geom_point() +
  labs(x = "Wage",
       y = "Education") +
  geom_smooth(method = "lm")

p4.2<-ggplot(d_missing, aes(log(wage), education)) +
  geom_point() +
  labs(x = "Log(Wage)",
       y = "Education") +
  geom_smooth(method = "lm")

p4.3<-ggplot(d_missing, aes(wage, experience)) +
  geom_point() +
  labs(x = "Wage",
       y = "Experience") +
  geom_smooth(method = "lm")

p4.4<-ggplot(d_missing, aes(log(wage), experience)) +
  geom_point() +
  labs(x = "Log(Wage)",
       y = "Experience") +
  geom_smooth(method = "lm")

p4.5<-ggplot(d_missing, aes(wage, raceColor)) +
  geom_point() +
  labs(x = "Wage",
       y = "Experience") +
  geom_smooth(method = "lm")

p4.6<-ggplot(d_missing, aes(log(wage), raceColor)) +
  geom_point() +
  labs(x = "Log(Wage)",
       y = "raceColor") +
  geom_smooth(method = "lm")

#create historgrams to try and idenfiy patterns
p4.7<-ggplot(d_missing, aes(x=education)) + geom_histogram(binwidth=2)
p4.8<-ggplot(d_missing, aes(x=experience)) + geom_histogram(binwidth=2)
p4.9<-ggplot(d_missing, aes(x=raceColor)) + geom_histogram(binwidth=.5)

#show the plots
multiplot(p4.7, p4.8, p4.9)
```

```r
multiplot(p4.1, p4.2, p4.3, p4.4, p4.5, p4.6, cols=2)
```

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

Ideally we would not want to just "throw away" these data points becuas they still contian some useful information. However, we need to be careful on how we handle the missing values.

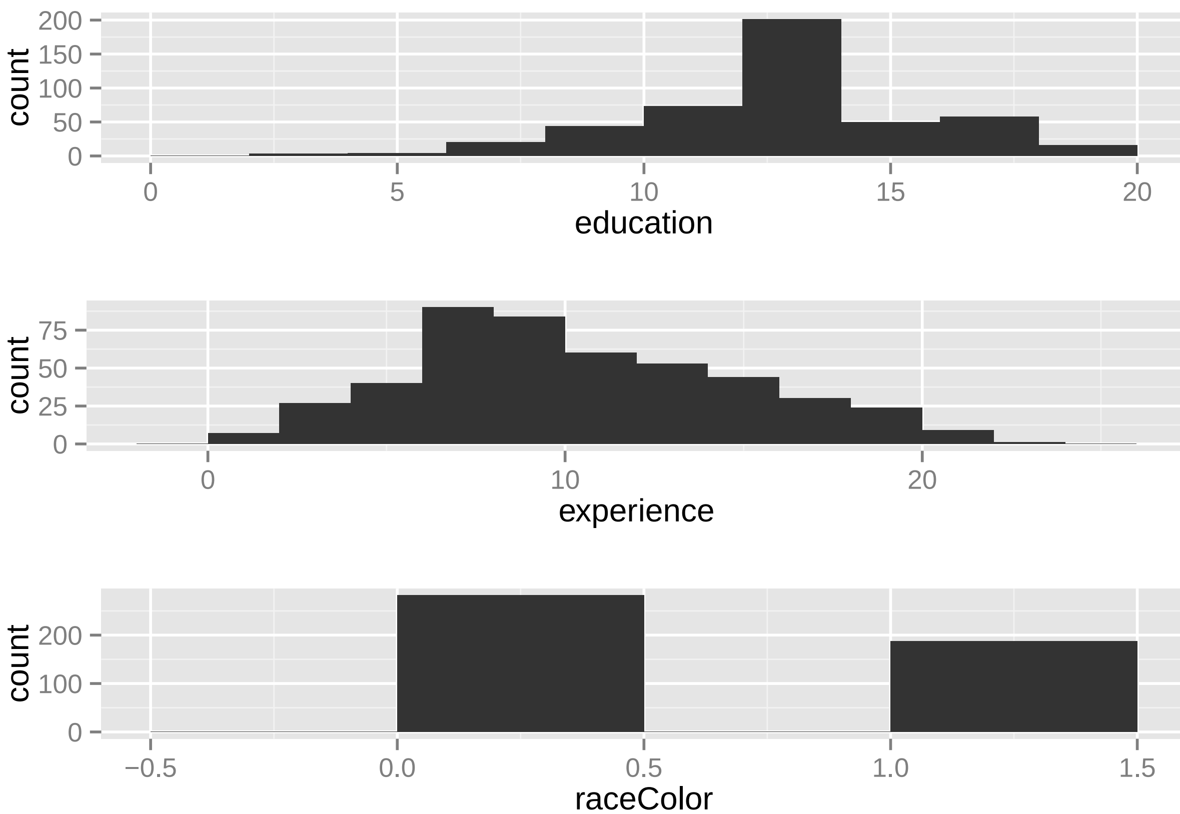

count
200
150
100
50
0
0
5
10
15
20
education
count
75
50
25
0
0
10
20
experience
count
200
100
0
−0.5
0.0
0.5
1.0
1.5
raceColor

Figure 8:

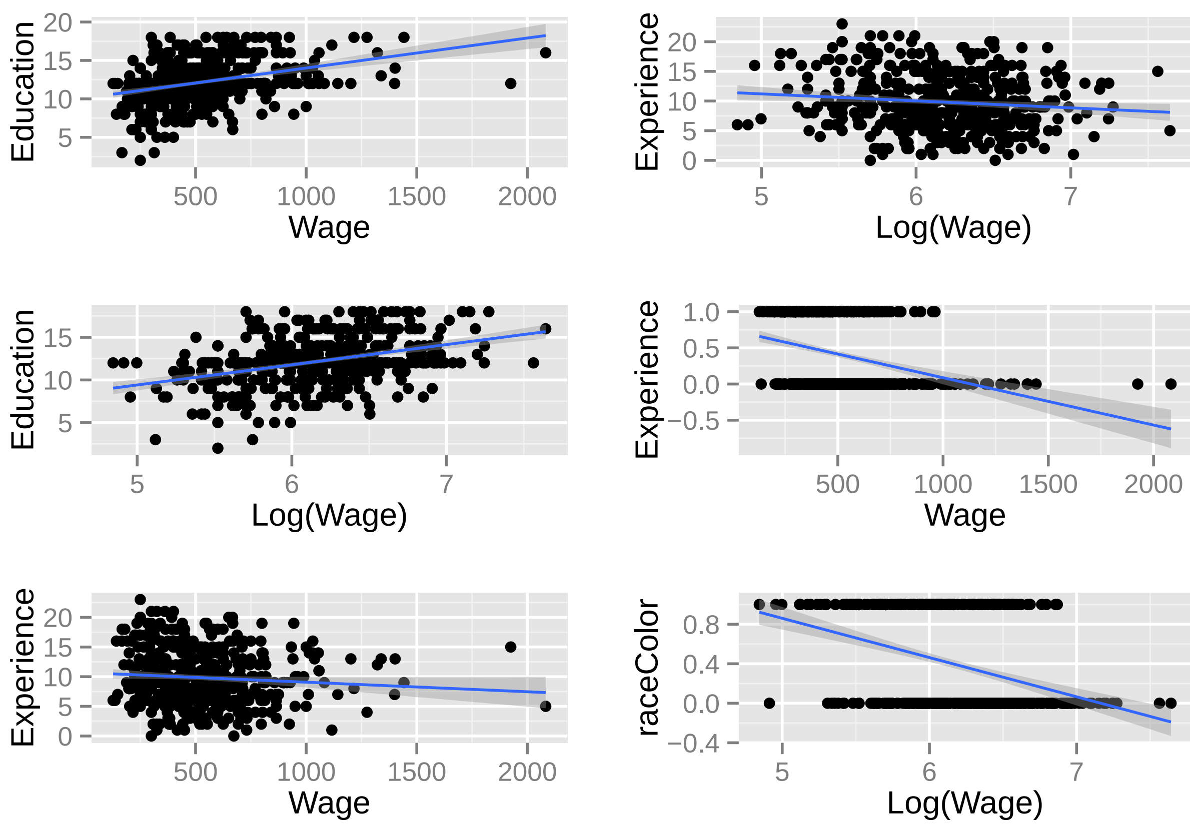

Education
Wage
Experience
Log(Wage)
Education
Log(Wage)
Experience
Wage
Experience
Wage
raceColor
Log(Wage)

Figure 9:

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?**

```
#create a copy of the dataset for this problem
d4.3<-d

#replace the missing values with the means of those values
d4.3$dad_education<-na.fill(d4.3$dad_education, mean(d4.3$dad_education, na.rm=T))
d4.3$mom_education<-na.fill(d4.3$mom_education, mean(d4.3$mom_education, na.rm=T))

#re-run the regression
model4.5.3<-lm(logWage~education + experience + experienceSquare + raceColor +
               dad_education + mom_education + rural + city, data=d4.3)
```

```
stargazer(model4.5, model4.5.3, type="latex", title="Question 4.5-3")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Mar 06, 2016 - 5:05:57 PM

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?**

```
d4<-d

model_dad<-lm(dad_education~education+experience+raceColor, data=d4)
model_mom<-lm(mom_education~education+experience+raceColor, data=d4)

coef_dad<-coef(model_dad)
coef_mom<-coef(model_mom)


for (i in 1:nrow(d)) {
  if (is.na(d$dad_education[i])==TRUE) {
    d$dad_education[i]= coef_dad[1]+coef_dad[2]*d$education[i]+
      coef_dad[3]*d$experience[i]+coef_dad[4]*d$raceColor[i]
  }
  if (is.na(d$mom_education[i])==TRUE) {
    d$mom_education[i]= coef_mom[1]+coef_mom[2]*d$education[i]+
      coef_mom[3]*d$experience[i]+coef_mom[4]*d$raceColor[i]
  }
}


#re-run the origional regression
model4.5.4<-lm(logWage~education + experience + experienceSquare + raceColor +
                 dad_education + mom_education + rural + city, data=d4)
```

```
#show the results of the the origional regression the final regression and the two sub regressions
stargazer(model4.5, model4.5.4, model_dad, model_mom, type="latex", title="Question 4.5-4", column.sep.w
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Mar 06, 2016 - 5:05:58 PM

Table 4: Question 4.5-3

	Dependent variable:	
	logWage	
	(1)	(2)
education	0.068^{***}	0.071^{***}
	(0.008)	(0.006)
experience	0.097^{***}	0.090^{***}
	(0.013)	(0.011)
experienceSquare	-0.003^{***}	-0.003^{***}
	(0.001)	(0.001)
raceColor	-0.213^{***}	-0.231^{***}
	(0.043)	(0.031)
dad_education	-0.001	-0.00004
	(0.005)	(0.004)
mom_education	0.011^{*}	0.003
	(0.006)	(0.005)
rural	-0.092^{***}	-0.095^{***}
	(0.031)	(0.026)
city	0.178^{***}	0.167^{***}
	(0.032)	(0.027)
Constant	4.642^{***}	4.729^{***}
	(0.141)	(0.123)
Observations	723	1,000
R^2	0.275	0.298
Adjusted R^2	0.267	0.292
Residual Std. Error	0.379 (df = 714)	0.376 (df = 991)
F Statistic	33.793^{***} (df = 8; 714)	52.617^{***} (df = 8; 991)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 5: Question 4.5-4

	Dependent variable:			
	logWage		dad_education	mom_education
	(1)	(2)	(3)	(4)
education	0.068***	0.068***	0.502***	0.433***
	(0.008)	(0.008)	(0.057)	(0.046)
experience	0.097***	0.097***	−0.148***	−0.077**
	(0.013)	(0.013)	(0.037)	(0.030)
experienceSquare	−0.003***	−0.003***		
	(0.001)	(0.001)		
raceColor	−0.213***	−0.213***	−2.121***	−1.468***
	(0.043)	(0.043)	(0.312)	(0.232)
dad_education	−0.001	−0.001		
	(0.005)	(0.005)		
mom_education	0.011*	0.011*		
	(0.006)	(0.006)		
rural	−0.092***	−0.092***		
	(0.031)	(0.031)		
city	0.178***	0.178***		
	(0.032)	(0.032)		
Constant	4.642***	4.642***	4.939***	5.593***
	(0.141)	(0.141)	(1.019)	(0.827)
Observations	723	723	761	872
R^2	0.275	0.275	0.309	0.274
Adjusted R^2	0.267	0.267	0.306	0.271
Residual Std. Error	0.379 (df = 714)	0.379 (df = 714)	3.122 (df = 757)	2.669 (df = 868)
F Statistic	33.793*** (df = 8; 714)	33.793*** (df = 8; 714)	112.815*** (df = 3; 757)	108.992*** (df = 3; 868)
Note:				*p<0.1; **p<0.05; ***p<0.01

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

```
stargazer(model4.5, model4.5.3, model4.5.4, type="latex", title="Question 4.5-5")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Mar 06, 2016 - 5:05:58 PM

Table 6: Question 4.5-5

	Dependent variable:		
	logWage		
	(1)	(2)	(3)
education	0.068***	0.071***	0.068***
	(0.008)	(0.006)	(0.008)
experience	0.097***	0.090***	0.097***
	(0.013)	(0.011)	(0.013)
experienceSquare	−0.003***	−0.003***	−0.003***
	(0.001)	(0.001)	(0.001)
raceColor	−0.213***	−0.231***	−0.213***
	(0.043)	(0.031)	(0.043)
dad_education	−0.001	−0.00004	−0.001
	(0.005)	(0.004)	(0.005)
mom_education	0.011*	0.003	0.011*
	(0.006)	(0.005)	(0.006)
rural	−0.092***	−0.095***	−0.092***
	(0.031)	(0.026)	(0.031)
city	0.178***	0.167***	0.178***
	(0.032)	(0.027)	(0.032)
Constant	4.642***	4.729***	4.642***
	(0.141)	(0.123)	(0.141)
Observations	723	1,000	723
R^2	0.275	0.298	0.275
Adjusted R^2	0.267	0.292	0.267
Residual Std. Error	0.379 (df = 714)	0.376 (df = 991)	0.379 (df = 714)
F Statistic	33.793*** (df = 8; 714)	52.617*** (df = 8; 991)	33.793*** (df = 8; 714)
Note:			*p<0.1; **p<0.05; ***p<0.01

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 $\mathbf{z_1}$ 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 $\mathbf{z_1}$ 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 $\mathbf{z_1}$ and $\mathbf{z_2}$ as instrument variables. Interpret the results. How does the coefficient estimate on education change?**

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, there is only one NA or missing value so we will omit that case. We then look for any potential outliers that we may want to omit or any anomalies in the data.

```
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
ggplot(data, aes(x=voteshare)) + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

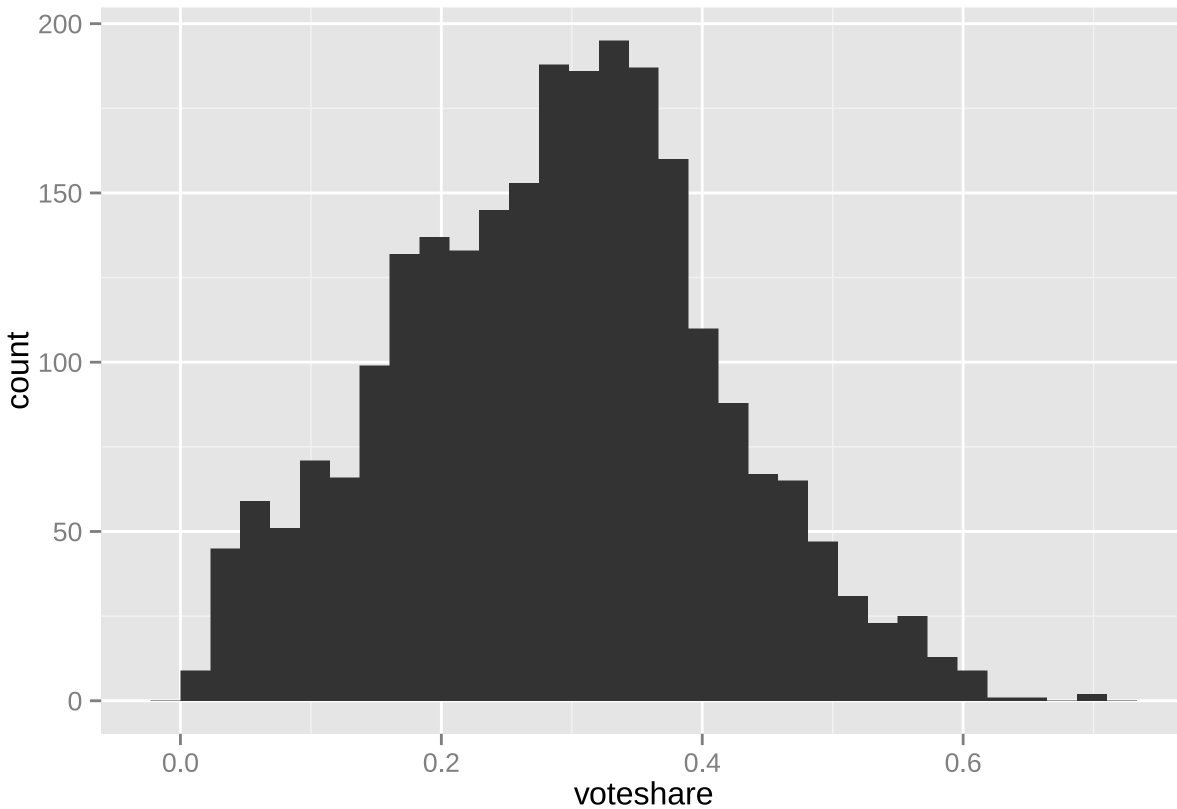


Figure 10:

```
ggplot(data, aes(x=absolute_wealth)) + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

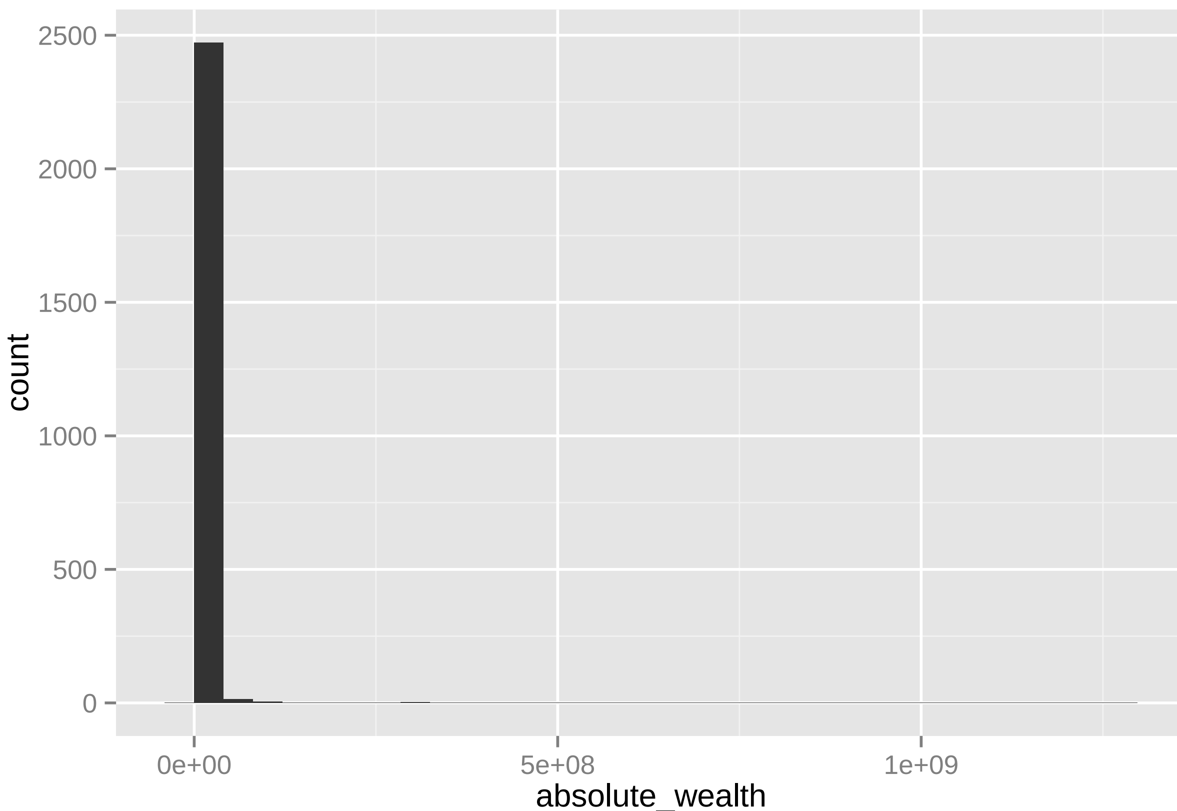


Figure 11:

```
#find any missing values
sum(is.na(data$voteshare))
```

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

```
sum(is.na(data$absolute_wealth))
```

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

```
sum(is.na(data$urb))
```

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

```
sum(is.na(data$lit))
```

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

```
sum(is.na(data$region))
```

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

```
#only take the complete cases
data<-data[complete.cases(data),]

#take the log of absoulte wealth to rescale the variable
data$logwealth<-log(data$absolute_wealth)
ggplot(data, aes(x=logwealth)) + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

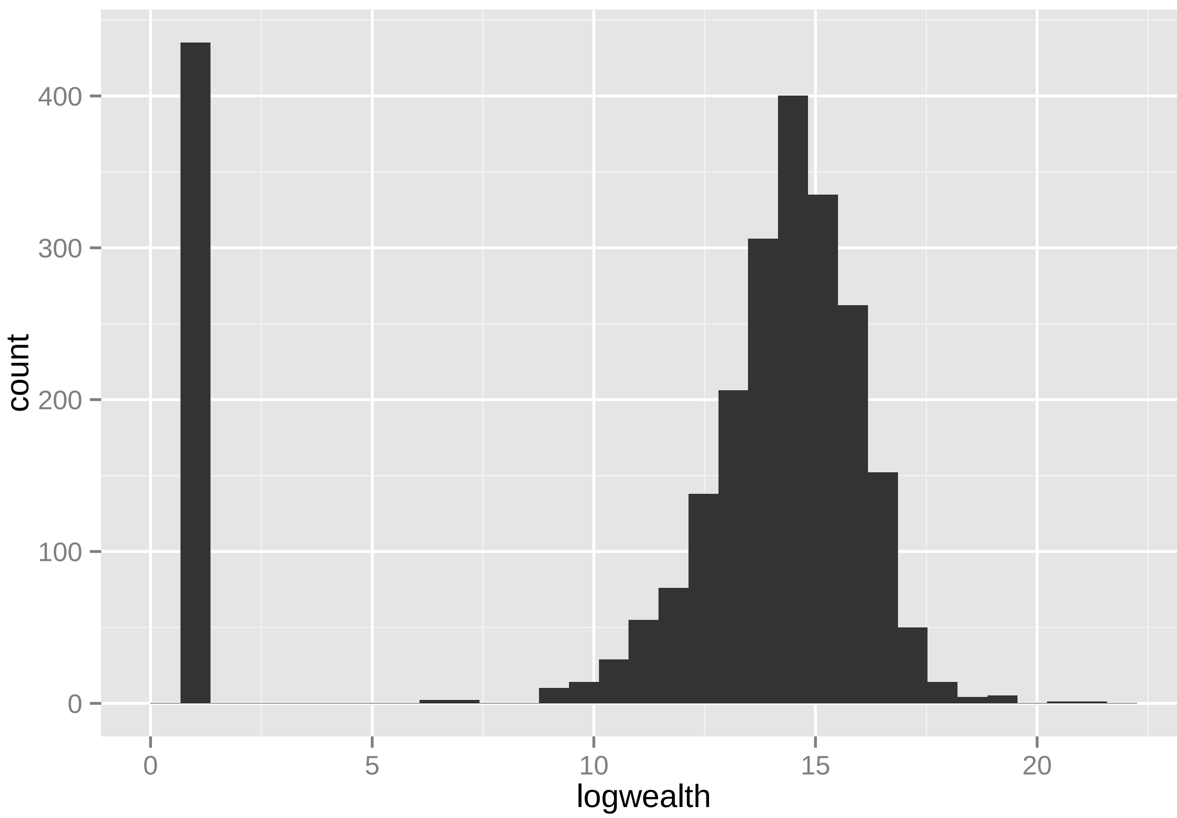


Figure 12:

From the histogram of logwealth we notice there is a large number of observations near zero with the same values. We look at a scaller plot of the voteshare and logwealth to see if any patterns are visable.

```
ggplot(data, aes(logwealth, voteshare)) + geom_point()
```

What we see is a lot of values with the exact same logwealth value. We do a little more analysis to try and indentify if there is a commonality among these points.

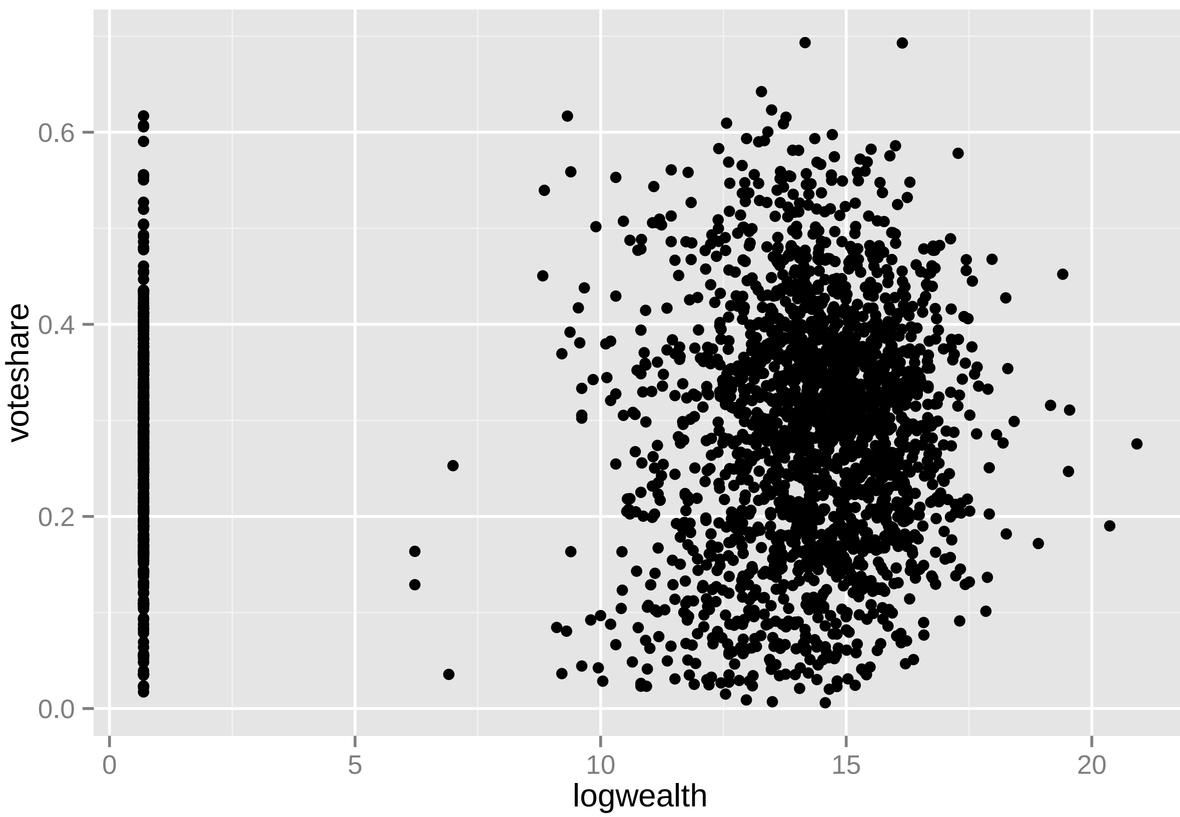


Figure 13:

```r
test<-subset(data, logwealth<1, )
head(test)
```

```
##     X   region        urb       lit voteshare absolute_wealth logwealth
## 22 22 Region 2 0.03219908 0.4358586 0.2953930               2 0.6931472
## 23 23 Region 2 0.03502421 0.3402133 0.3627455               2 0.6931472
## 41 41 Region 2 0.13326013 0.4918037 0.4546827               2 0.6931472
## 50 50 Region 2 0.07662392 0.3752757 0.3509711               2 0.6931472
## 56 56 Region 2 0.13676432 0.4046299 0.2938662               2 0.6931472
## 91 91 Region 2 0.08724256 0.2751910 0.2506499               2 0.6931472
```

```r
ggplot(test, aes(voteshare)) + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

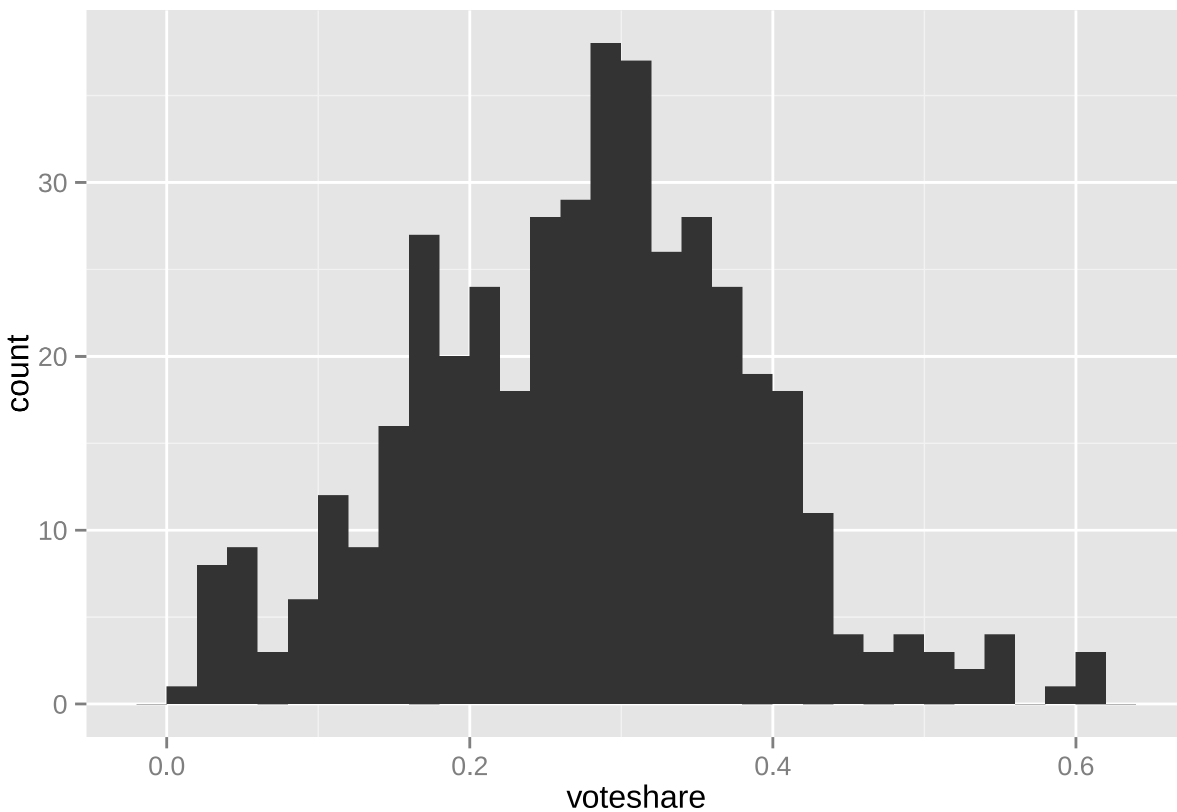


Figure 14:

```r
ggplot(test, aes(region)) + geom_histogram()
```

```r
ggplot(test, aes(urb)) + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

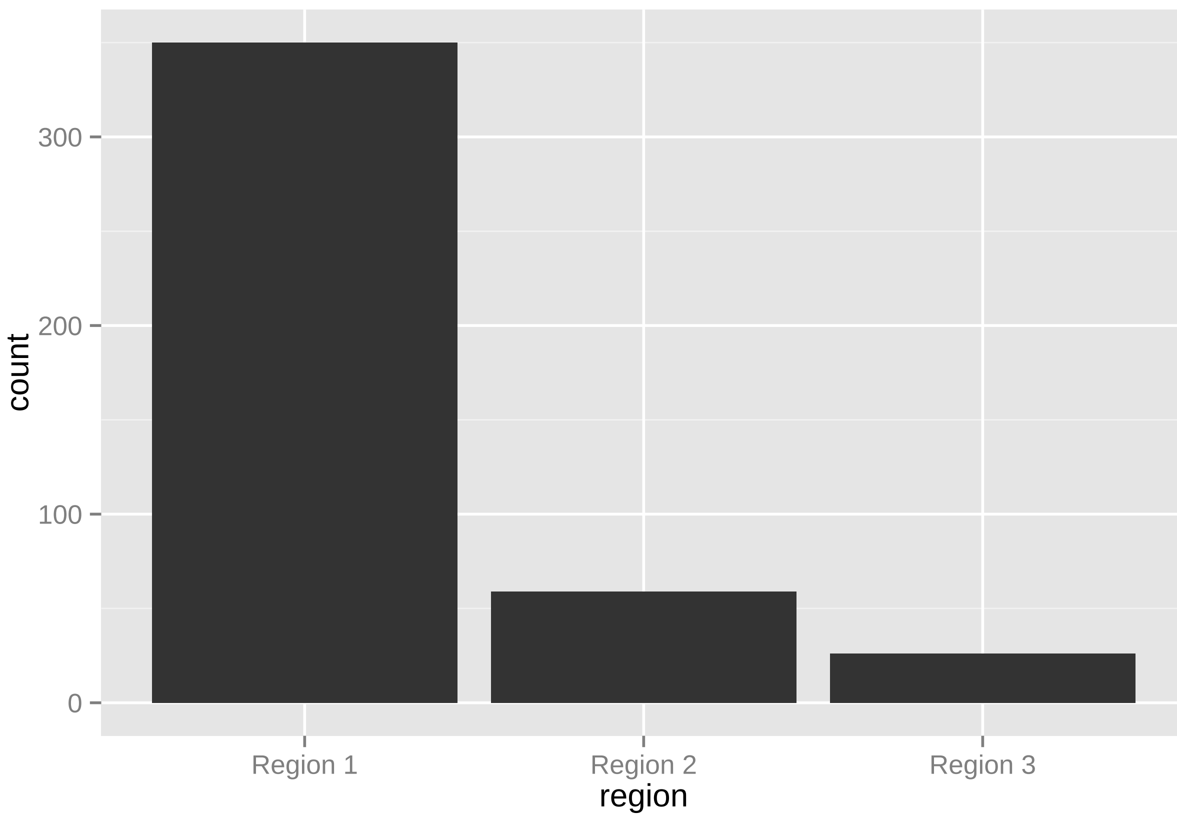


Figure 15:

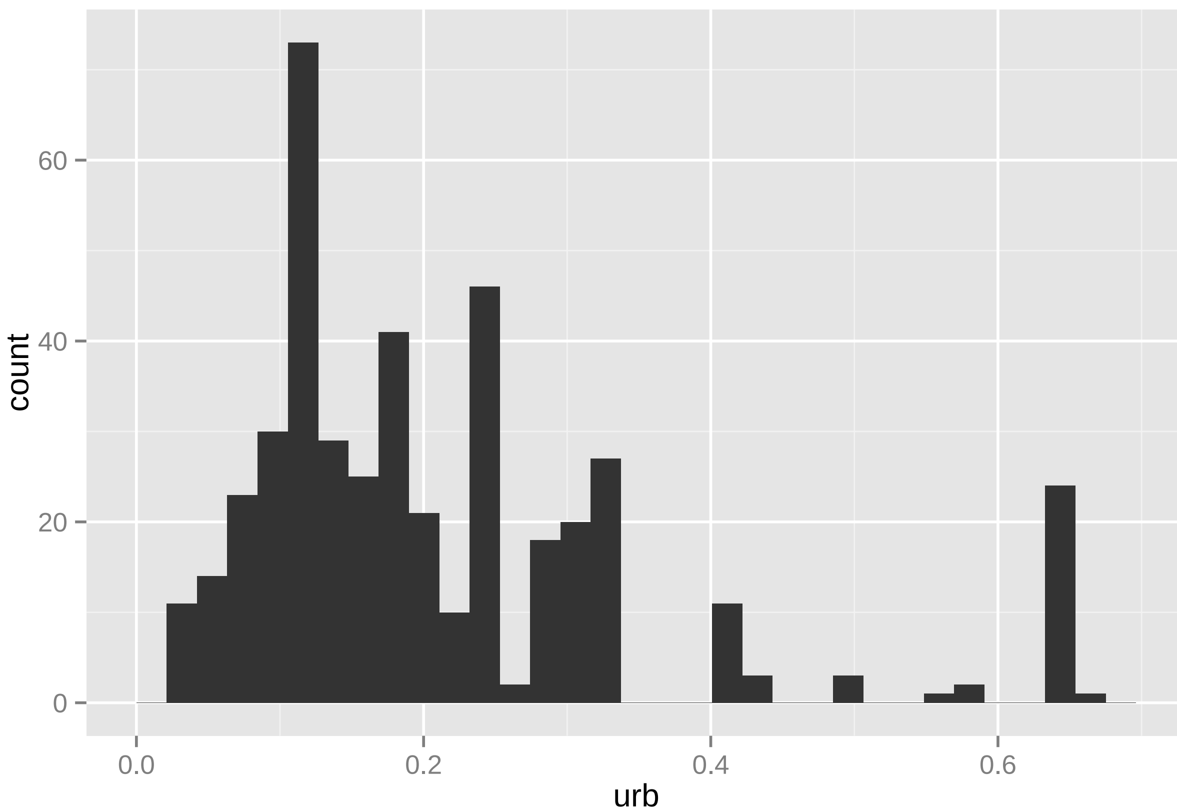


Figure 16:

```
ggplot(test, aes(lit)) + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

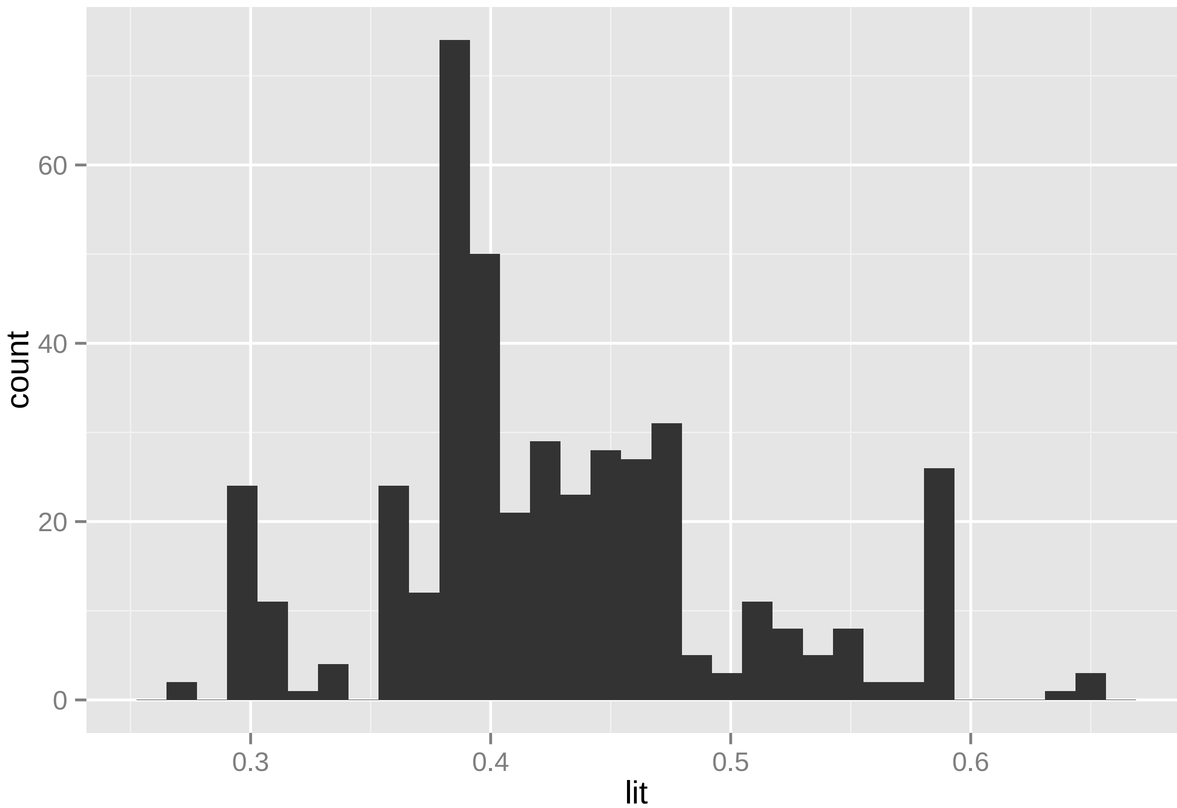


Figure 17:

Based on this analysis we can see that most of these identical values are from region 1. At this point we could omit these datapoints if we had a reason that they were inaccruate, but in continuing we will leave them in because it could be some underlying factor in region 1 that impacts this value and could be important to the end results.

Now to build a parsimonious model to try and answer the question of is wealth an electoral advantage, we will start off with the simple regression of just logweath on voteshare. We choose this model as a basic model because the other variables of urb and lit could have counding factors with wealth.

```
#run the basic model
model1<-lm(voteshare~logwealth, data=data)
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: Sun, Mar 06, 2016 - 5:05:59 PM

Based on the output results the model does show statistically significant results, however the R^2 is almost zero which indicates that the model does not fit the data very well at all. We would say there is not a linear relationship between wealth and voteshare.

Table 7: Question 5.1

	Dependent variable:
	voteshare
logwealth	0.001***
	(0.0005)
Constant	0.272***
	(0.006)
Observations	2,497
R^2	0.003
Adjusted R^2	0.003
Residual Std. Error	0.123 (df = 2495)
F Statistic	7.983*** (df = 1; 2495)
Note:	*p<0.1; **p<0.05; ***p<0.01

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?**

Based on the above results, it is unlikley that the square term will make much difference.

```
#run the model with the squared term added in
model2<-lm(voteshare~logwealth+ logwealth*logwealth, data=data)
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: Sun, Mar 06, 2016 - 5:05:59 PM

Table 8: Question 5.2

	Dependent variable:	
	voteshare	
	(1)	(2)
logwealth	0.001***	0.001***
	(0.0005)	(0.0005)
Constant	0.272***	0.272***
	(0.006)	(0.006)
Observations	2,497	2,497
R^2	0.003	0.003
Adjusted R^2	0.003	0.003
Residual Std. Error (df = 2495)	0.123	0.123
F Statistic (df = 1; 2495)	7.983***	7.983***
Note:	*p<0.1; **p<0.05; ***p<0.01	

The squared term does not dramatically change the model indicating wealthier candidates do not necessarily fare better in elections.

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=data)

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: Sun, Mar 06, 2016 - 5:05:59 PM

Table 9: Question 5.3

	Dependent variable:		
	voteshare		
	(1)	(2)	(3)
logwealth	0.001***	0.001***	0.001*
	(0.0005)	(0.0005)	(0.0005)
Constant	0.272***	0.272***	0.264***
	(0.006)	(0.006)	(0.006)
Region Fixed effects	No	No	Yes
Observations	2,497	2,497	2,497
R^2	0.003	0.003	0.018
Adjusted R^2	0.003	0.003	0.017
Residual Std. Error	0.123 (df = 2495)	0.123 (df = 2495)	0.122 (df = 2493)
F Statistic	7.983*** (df = 1; 2495)	7.983*** (df = 1; 2495)	15.255*** (df = 3; 2493)
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?**

It is unlikely that we have a causal and unbiased estimate because there are likely omitted variables in our linear model.

To have an unbiased and causal estimate we would first need to statisfy the assumptions for OLS which are: 1. (the model is) linear in parameters 2. random sampling 3. no perfect collinearity (among the independent variables) 3. zero mean (of the errors) and zero correlation (with any of the independent variables)

Additionally, causality is about whether manipulations to the independent variable influence the dependent variable but not the error term. For a model to be causal, we need to be able to introduce a manipulation in x, dx, that (we expect) will cause a change in y, dy, while the error term u (that includes both the idiosyncratic error and the individual time-constant or fixed effect) needs to stay unchanged as we manipulate x. I.e., as long as

$$\frac{\partial u}{\partial x} = 0$$

we can claim that the effect of x is

$$\frac{\partial y}{\partial x} = \beta_1.$$

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?**

In order to run a difference-in-difference design we need to have a time component. In this situation if we had data from two different election years we would be able to run a difference-in-difference design. The equation would be similar to the one shown below where we take the difference in the data between the two measurements and the difference in the error and run the regression model on the differenced equation.

$$\Delta voteshare = \delta_0 + \beta_1 \Delta logwealth + \Delta u$$

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.**