W271-2 – Spring 2016 – **Lab 2**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

March 7, 2016

Contents

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, $\mathbf{E(Y|X)}$.**

Note that for a continuous random variable, Z, the probability density function is:

$$f_Z(z) = \begin{cases} \dfrac{1}{b-a} & a \leq z \leq \ b \\ 0 & \text{otherwise} \end{cases}$$

In the case of variable X this gives:

$E(X) = \displaystyle\int_{x=0}^{1} x \cdot dx = \frac{1^2}{2} = \frac{1}{2}$

In the case of variable Y, note that the length of the ruler is simply X

$\mathbf{E(Y|X)} = \mathbf{\dfrac{X}{2}}$

2. **Find the unconditional expectation of Y. One way to do this is to apply the law of iterated expectations, which states that $\mathbf{E(Y) = E(E(Y|X))}$. 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.**

$$\mathbf{E(Y)} = E(E(Y|X)) = \int_X E(Y|X) \cdot f_X(x) \cdot dx = \int_{x=0}^{1} \frac{x}{2} \cdot 1 cdot dx = \frac{1}{2}\int_{x=0}^{1} x = \mathbf{\frac{1}{4}}$$

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

Using the definition from part 1:

$$f_{X,Y}(x, y) = \begin{cases} \dfrac{1}{x} & \mathbf{x \in (0, 1), y \in (0, x)} \\ 0 & \text{otherwise} \end{cases}$$

4. **Find the conditional probability density function of X given Y, $\mathbf{f_{X|Y}}$.**
5. **Find the expectation of X, given that Y is 1/2, $\mathbf{E(X|Y = 1/2)}$.**

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, $\mathbf{l}(\theta)$ in terms of $\mathbf{y_1, y_2, \ldots, y_n}$.**

The pdf of a given observation, y, is as follows:

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

And the liklihood function has the form

$$L(\theta) = \begin{cases} \frac{1}{\theta^n} & 0 \leq x_i \leq \theta (i = 1, \ldots, n) \\ 0 & \text{otherwise} \end{cases}$$

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

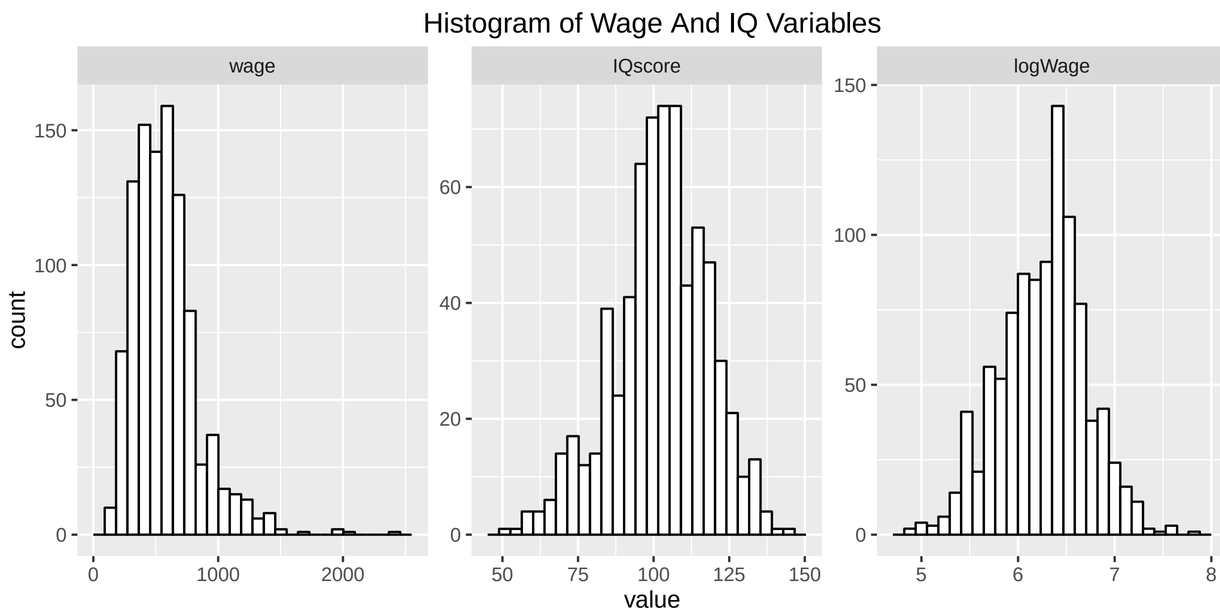
Question 4.1

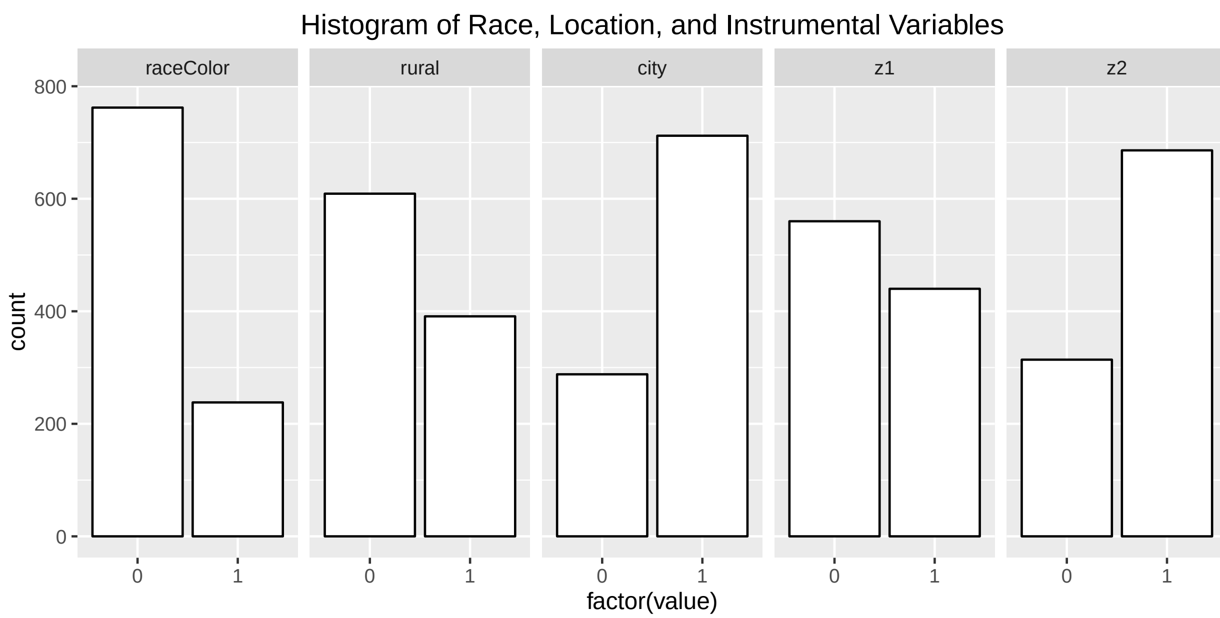
Conduct an univariate analysis (using tables, graphs, and descriptive statistics found in the last 7 lectures) of all of the variables in the dataset.

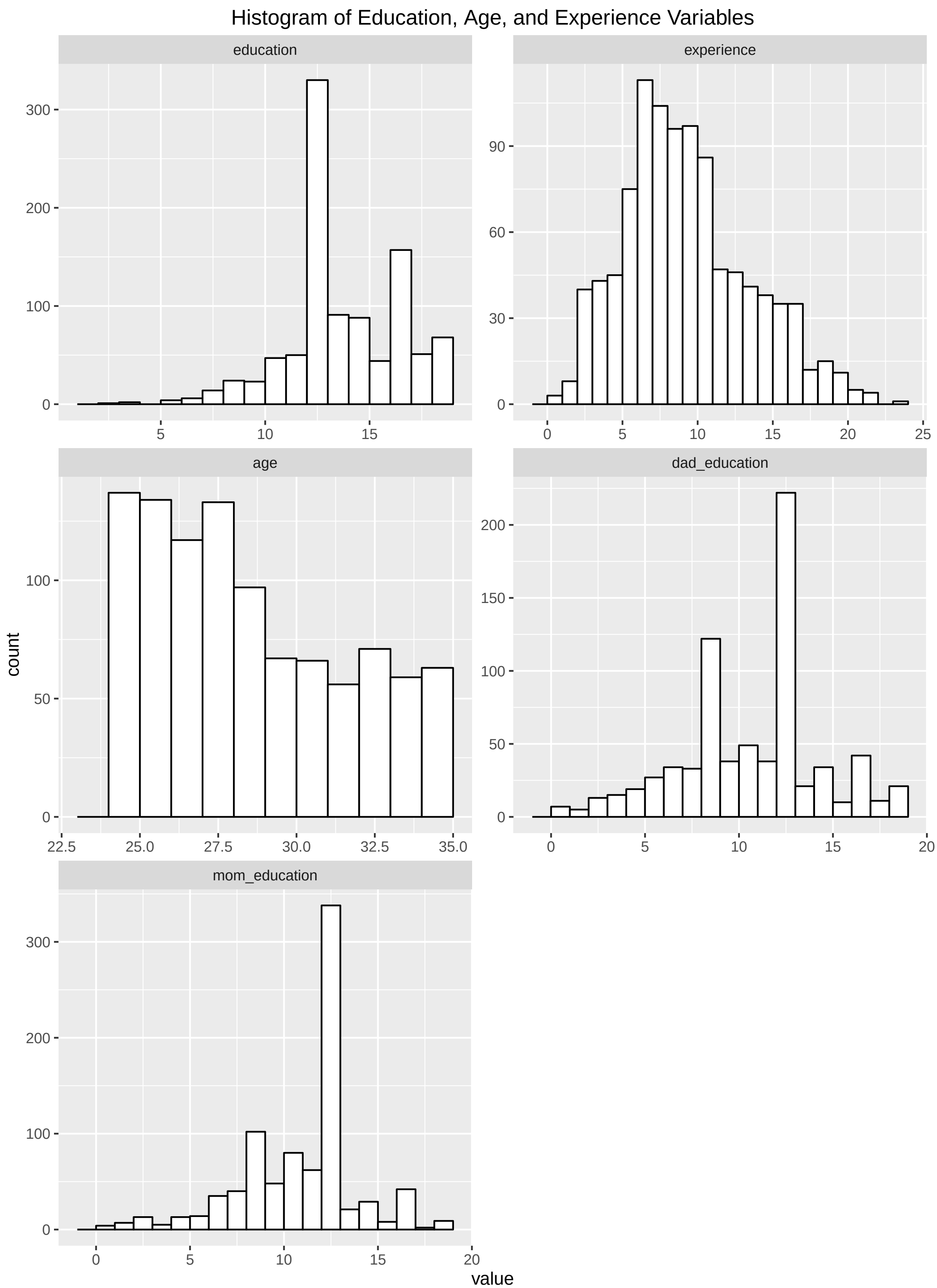
Also, create two variables: (1) natural log of wage (name it `logWage`) (2) square of experience (name it `experienceSquare`)

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
X	1,000	1,466.678	866.535	5	715.5	1,431.5	2,212	3,009
wage	1,000	578.783	266.569	127	400	543	702.5	2,404
education	1,000	13.219	2.729	2	12	12	16	18
experience	1,000	8.788	4.221	0	6	8	11	23
age	1,000	28.007	3.118	24	25	27	30	34
raceColor	1,000	0.238	0.426	0	0	0	0	1
dad_education	761	10.181	3.748	0	8	11	12	18
mom_education	872	10.451	3.126	0	8	12	12	18
rural	1,000	0.391	0.488	0	0	0	1	1
city	1,000	0.712	0.453	0	0	1	1	1
z1	1,000	0.440	0.497	0	0	0	1	1
z2	1,000	0.686	0.464	0	0	1	1	1
IQscore	684	102.273	15.843	50	93	103	113	144
logWage	1,000	6.263	0.447	4.844	5.991	6.297	6.555	7.785
experienceSquare	1,000	95.030	86.786	0	36	64	121	529


Histogram of Wage And IQ Variables
wage
IQscore
logWage
count
value


Histogram of Race, Location, and Instrumental Variables
raceColor
rural
city
z1
z2
count
factor(value)


Histogram of Education, Age, and Experience Variables
education
experience
age
dad_education
mom_education
count
value

Question 4.2

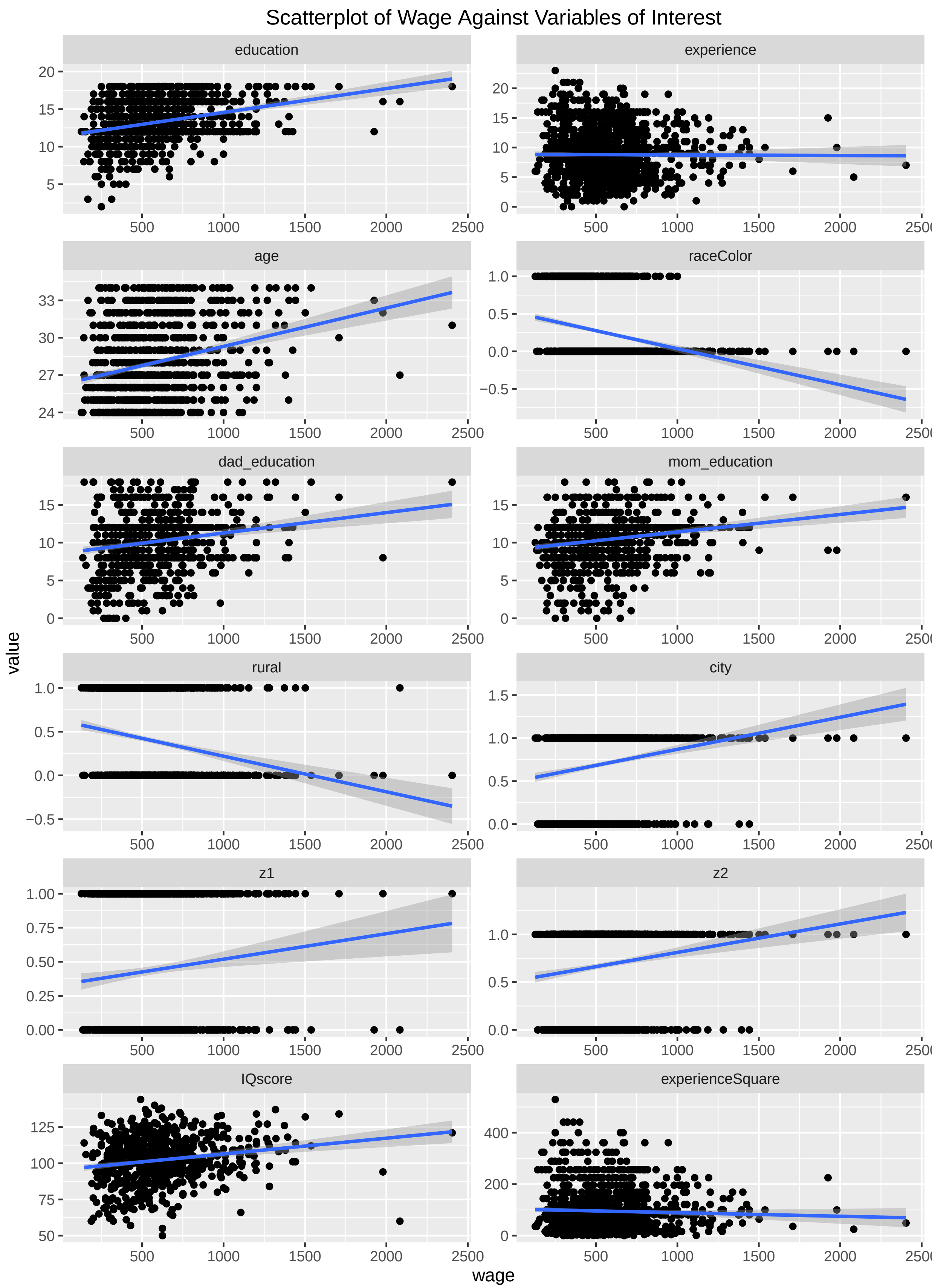
Conduct a bivariate analysis (using tables, graphs, descriptive statistics found in the last 7 lectures) of `wage` and `logWage` and all the other variables in the datasets.

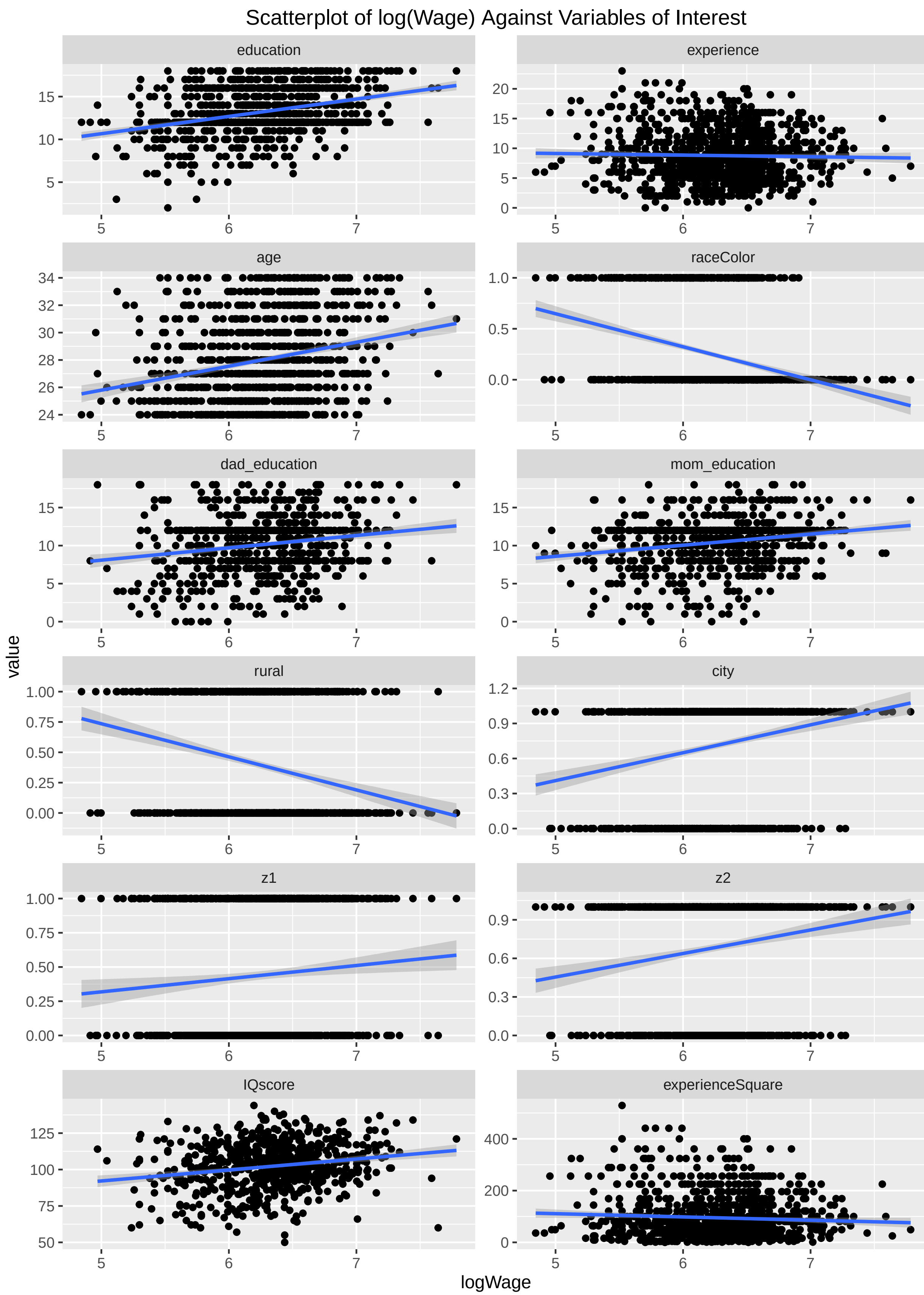
Table 2: Correlations for Wage

	Wage
education	0.310
experience	-0.006
age	0.264
raceColor	-0.301
dad_education	0.190
mom_education	0.198
rural	-0.222
city	0.220
z1	0.101
z2	0.171
IQscore	0.186
logWage	0.946
experienceSquare	-0.043

Table 3: Correlations for log(Wage)

	log(Wage)
wage	0.946
education	0.332
experience	-0.029
age	0.251
raceColor	-0.341
dad_education	0.189
mom_education	0.210
rural	-0.250
city	0.236
z1	0.087
z2	0.177
IQscore	0.201
experienceSquare	-0.065


Scatterplot of Wage Against Variables of Interest
education
experience
age
raceColor
dad_education
mom_education
rural
city
z1
z2
IQscore
experienceSquare
value
wage



Scatterplot of log(Wage) Against Variables of Interest
education
experience
age
raceColor
dad_education
mom_education
rural
city
z1
z2
IQscore
experienceSquare
value
logWage

Question 4.3

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

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

Table 4: Regression Summary

	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

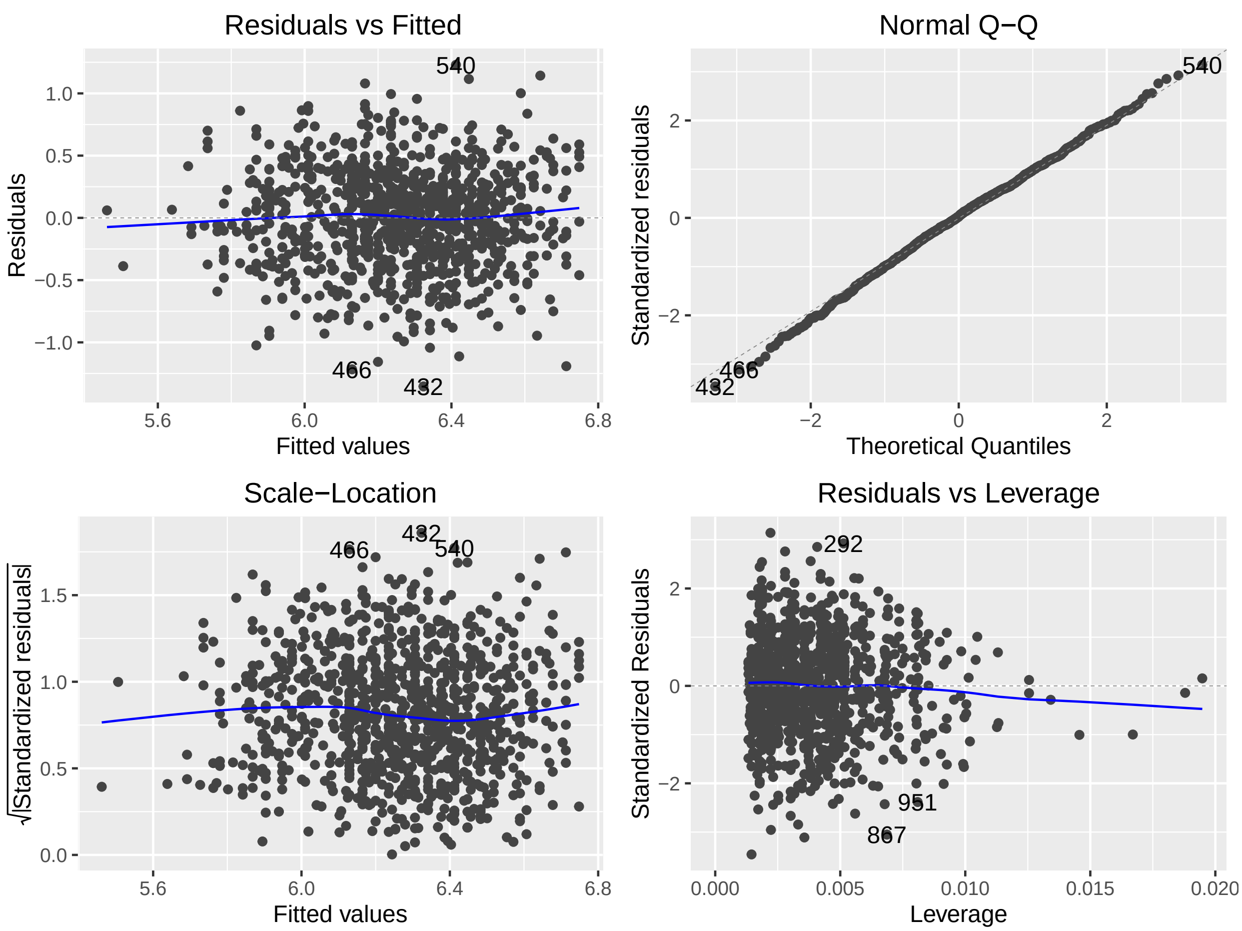
2. **Explain why the degrees of freedom takes on the specific value you observe in the regression output.**

The overall degrees of freedom (F(3, 996) is one smaller than we might expect otherwise. This is because the paramter for age is a linearly dependent combination of the other parameters and it's coefficient cannot be estimated. R automatically removes this variable from the model, and thus our regression output reflects the degrees of freedom for a model with 3 parameters rather than 4. Thus estimating this model is equivalent to estimating $log(wage) = education + experience + raceColor$.

Table 5: Regression summary

	Dependent variable:
	log(Wages)
Education	0.080***
	(0.006)
Experience	0.035***
	(0.004)
Race (White or Non-white)	−0.261***
	(0.030)
Constant	4.962***
	(0.115)
F Statistic	100.278***
df	3; 996
Observations	1,000
R^2	0.236
Adjusted R^2	0.234
Residual Std. Error	0.392

·p<0.1; *p<0.05; **p<0.01; ***p<0.001



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 inability to estimate the age coefficient is an unexpected result. Looking at the corelation matrix from above, we don't see that any variable is perfectly correlated with $log(wage)$. However, if we take the correlation of *age* with the sum of *education* and *experience*, we see that *age* is perfectly correlated with the two variables. Since the model already incorporates all of the information contained in the age variable, we can simply remove it from the model and not lose anything.

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

Holding other covariates constant, a one year increase in *education* was associated with a statistically significant increase in $log(wage)$, (β = 0.080, std. error = 0.006, t = 12.394, p < .001). This effect also represents a practically significant wage return on education.

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

Holding other covariates constant, a one year increase in *experience* was associated with a statistically significant increase in $log(wage)$, (β = 0.035, std. error = 0.004, t = 8.81, p < .001). This effect also represents a practically significant wage return on experience.

Question 4.4

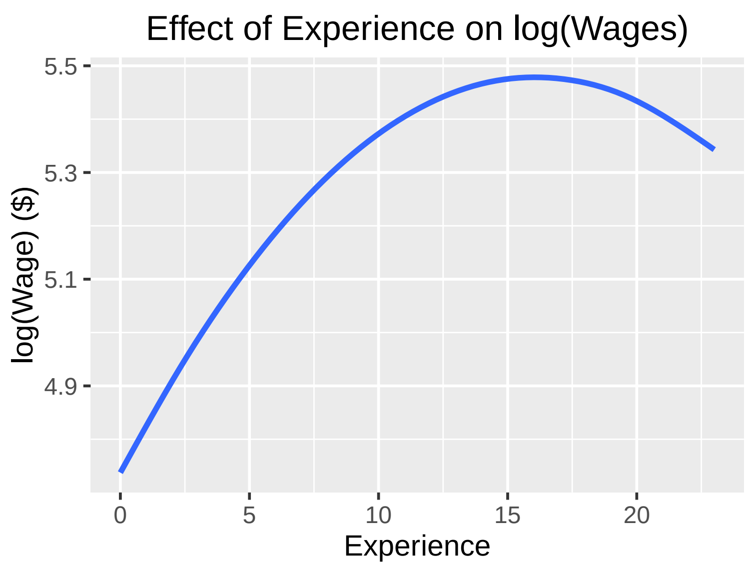
Regress *log(wage)* on education, experience, experienceSquare, and race-Color.

Table 6: Regression summary

	Dependent variable:
	log(Wages)
Education	0.079***
	(0.006)
Experience	0.092***
	(0.011)
Expereince^2	−0.003***
	(0.001)
Race (White or Non-white)	−0.263***
	(0.030)
Constant	4.736***
	(0.120)
F Statistic	84.960***
df	4; 995
Observations	1,000
R^2	0.257
Adjusted R^2	0.254
Residual Std. Error	0.387

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

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?**

The estimated effect of experience at 10 years on wages is $10 \cdot 0.092 + 10\hat{}2 \cdot$ -0.003 = 0.637

Question 4.5

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

Table 7: Regression summary

	Dependent variable:
	log(Wages)
Education	0.068***
	(0.008)
Experience	0.097***
	(0.013)
Expereince^2	−0.003***
	(0.001)
Race (White or Non-white)	−0.213***
	(0.041)
Father's Education	−0.001
	(0.006)
Mother's Education	0.011
	(0.007)
Rural (Yes or No)	−0.092**
	(0.032)
City (Yes or No)	0.178***
	(0.032)
Constant	4.642***
	(0.150)
F Statistic	32.592***
df	8; 714
Observations	723
R^2	0.275
Adjusted R^2	0.267
Residual Std. Error	0.379

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

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 723 observations in the regression. A comparision of the original data and the observations in the regression reveals that the missing observations tend to have lower wages and education, and are less likely to live in the city. The missing observations have more experienceand, and are more likely to be non-white and live in rural areas.

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

The clear pattern of missing values representing more rural, non-white observations with lower education indicates that these values should not simply be discarded from the model. In order to accurately assess the effect of covariates these values are important for the regression.

Table 8: All Observations

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
X	1,000	1,466.678	866.535	5	715.5	1,431.5	2,212	3,009
wage	1,000	578.783	266.569	127	400	543	702.5	2,404
education	1,000	13.219	2.729	2	12	12	16	18
experience	1,000	8.788	4.221	0	6	8	11	23
age	1,000	28.007	3.118	24	25	27	30	34
raceColor	1,000	0.238	0.426	0	0	0	0	1
dad_education	761	10.181	3.748	0	8	11	12	18
mom_education	872	10.451	3.126	0	8	12	12	18
rural	1,000	0.391	0.488	0	0	0	1	1
city	1,000	0.712	0.453	0	0	1	1	1
z1	1,000	0.440	0.497	0	0	0	1	1
z2	1,000	0.686	0.464	0	0	1	1	1
IQscore	684	102.273	15.843	50	93	103	113	144
logWage	1,000	6.263	0.447	4.844	5.991	6.297	6.555	7.785
experienceSquare	1,000	95.030	86.786	0	36	64	121	529

Table 9: Missing Father's Education

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
X	239	1,675.259	864.761	15	921.5	1,712	2,515.5	3,009
wage	239	529.531	267.216	127	351.5	465	640.5	2,083
education	239	12.121	2.696	3	11	12	13	18
experience	239	10.527	4.223	1	7	10	14	23
age	239	28.649	3.090	24	26	28	31	34
raceColor	239	0.464	0.500	0	0	0	1	1
mom_education	149	8.987	3.397	0	7	9	12	18
rural	239	0.540	0.499	0	0	1	1	1
city	239	0.661	0.474	0	0	1	1	1
z1	239	0.397	0.490	0	0	0	1	1
z2	239	0.669	0.471	0	0	1	1	1
IQscore	132	96.038	17.421	50	85	98.5	107	135
logWage	239	6.163	0.465	4.844	5.862	6.142	6.462	7.642
experienceSquare	239	128.577	98.151	1	49	100	196	529

Table 10: Missing Mother's Education

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
X	128	1,765.523	870.253	53	996	1,796	2,572	3,009
wage	128	525.938	252.627	142	371.5	492.5	625	2,083
education	128	11.891	2.713	2	10	12	13	18
experience	128	10.898	4.742	0	7	10	15	21
age	128	28.789	3.513	24	26	28	32	34
raceColor	128	0.461	0.500	0	0	0	1	1
dad_education	38	9.184	3.840	2	6	9.5	12	16
rural	128	0.523	0.501	0	0	1	1	1
city	128	0.656	0.477	0	0	1	1	1
z1	128	0.414	0.494	0	0	0	1	1
z2	128	0.680	0.468	0	0	1	1	1
IQscore	71	93.352	15.900	60	80.5	96	104.5	124
logWage	128	6.175	0.418	4.956	5.918	6.199	6.438	7.642
experienceSquare	128	141.086	108.986	0	49	100	225	441

Table 11: Actual Observations

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
X	723	1,399.028	853.489	5	680.5	1,314	2,125	2,998
wage	723	597.080	268.094	136	409	570	721	2,404
education	723	13.651	2.616	3	12	13	16	18
experience	723	8.145	4.005	0	5	8	10	21
age	723	27.797	3.066	24	25	27	30	34
raceColor	723	0.158	0.365	0	0	0	0	1
dad_education	723	10.234	3.739	0	8	11	12	18
mom_education	723	10.752	2.982	0	9	12	12	18
rural	723	0.346	0.476	0	0	0	1	1
city	723	0.730	0.444	0	0	1	1	1
z1	723	0.448	0.498	0	0	0	1	1
z2	723	0.685	0.465	0	0	1	1	1
IQscore	531	104.081	15.062	60	95	105	114	144
logWage	723	6.297	0.442	4.913	6.014	6.346	6.581	7.785
experienceSquare	723	82.367	78.067	0	25	64	100	441

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?**

```
data7 <- data
data7$dad_education <- na.fill(data7$dad_education, mean(data7$dad_education, na.rm=T))
data7$mom_education <- na.fill(data7$mom_education, mean(data7$mom_education, na.rm=T))
model6 <- lm(logWage ~ education + experience + experienceSquare + raceColor +
               dad_education + mom_education + rural + city, data = data7)
```

Table 12: Regression summary

	Dependent variable:
	log(Wages)
Education	0.071***
	(0.007)
Experience	0.090***
	(0.011)
Expereince^2	−0.003***
	(0.001)
Race (White or Non-white)	−0.231***
	(0.031)
Father's Education	−0.00004
	(0.005)
Mother's Education	0.003
	(0.005)
Rural (Yes or No)	−0.095***
	(0.027)
City (Yes or No)	0.167***
	(0.026)
Constant	4.729***
	(0.128)
F Statistic	54.109***
df	8; 991
Observations	1,000
R^2	0.298
Adjusted R^2	0.292
Residual Std. Error	0.376

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

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?**

```
data8 <-data
model_dad<-lm(dad_education~education+experience+raceColor, data=data8)
model_mom<-lm(mom_education~education+experience+raceColor, data=data8)
coef_dad<-coef(model_dad)
coef_mom<-coef(model_mom)
# Impute values for missing observations
for (i in 1:nrow(data8)) {
```

```
  if (is.na(data8$dad_education[i])==TRUE) {
    data8$dad_education[i]= coef_dad[1]+coef_dad[2]*data8$education[i]+
      coef_dad[3]*data8$experience[i]+coef_dad[4]*data8$raceColor[i]
  }
  if (is.na(data8$mom_education[i])==TRUE) {
    data8$mom_education[i]= coef_mom[1]+coef_mom[2]*data8$education[i]+
      coef_mom[3]*data8$experience[i]+coef_mom[4]*data8$raceColor[i]
  }
}
```

Table 13: Regression summary

	Dependent variable:
	log(Wages)
Education	0.070***
	(0.007)
Experience	0.089***
	(0.011)
Expereince^2	−0.003***
	(0.001)
Race (White or Non-white)	−0.227***
	(0.031)
Father's Education	0.002
	(0.005)
Mother's Education	0.002
	(0.006)
Rural (Yes or No)	−0.095***
	(0.027)
City (Yes or No)	0.167***
	(0.026)
Constant	4.727***
	(0.125)
F Statistic	54.147***
df	8; 991
Observations	1,000
R^2	0.298
Adjusted R^2	0.293
Residual Std. Error	0.376

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

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

Table 14:

	Dependent variable:		
	log(Wages)		
	(1)	(2)	(3)
Education	0.068***	0.071***	0.070***
	(0.008)	(0.007)	(0.007)
Experience	0.097***	0.090***	0.089***
	(0.013)	(0.011)	(0.011)
Expereince^2	−0.003***	−0.003***	−0.003***
	(0.001)	(0.001)	(0.001)
Race (White or Non-white)	−0.213***	−0.231***	−0.227***
	(0.041)	(0.031)	(0.031)
Father's Education	−0.001	−0.00004	0.002
	(0.006)	(0.005)	(0.005)
Mother's Education	0.011	0.003	0.002
	(0.007)	(0.005)	(0.006)
Rural (Yes or No)	−0.092**	−0.095***	−0.095***
	(0.032)	(0.027)	(0.027)
City (Yes or No)	0.178***	0.167***	0.167***
	(0.032)	(0.026)	(0.026)
Constant	4.642***	4.729***	4.727***
	(0.150)	(0.128)	(0.125)
F Statistic	32.592***	54.109***	54.147***
df	8; 714	8; 991	8; 991
Observations	723	1,000	1,000
R^2	0.275	0.298	0.298
Adjusted R^2	0.267	0.292	0.293
Residual Std. Error	0.379	0.376	0.376

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

Comparing the regression results, it's worth noting that the magnitude and significance of the coefficients are broadly similar. However, given that the values that we imputed are for covariates that are neither statistically nor practically significant, I would prefer a model that used all of the observations and did not include the covariates for parent's education.

Question 4.6

1. **Consider using $\mathbf{z_1}$ as the instrumental variable (IV) for education. What assumptions are needed on $\mathbf{z_1}$ and the error term (call it, u)?**

Formally, we need to assume that the instrumental variable is uncorrelated with the error term. Ie.

$$Cov(z_{1j}, u) = 0 \; \forall j = 1, 2, \ldots, k$$

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?**

Yes. Living in an area with recent education policy changes could correlate to a number of variables that relate to the earnings in that area, such as a person's trust in educational and governmental institutions, or the willingness of people to pay taxes to support education spending. As with most instrumental variables, the assumption that it is truly uncorrelated with the error term can always be questioned, although it may represent an improvement in estimating coefficients over straightforward OLS.

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_stage_a <- lm(education ~ z1, data = data )
first_stage_b <- lm(education ~ z2, data = data)
second_stage_a <- lm(data$logWage ~ first_stage_a$fitted + data$experience +
                     data$experienceSquare + data$raceColor + data$dad_education +
                     data$mom_education + data$rural + data$city)
second_stage_b <- lm(data$logWage ~ first_stage_b$fitted + data$experience +
                       data$experienceSquare + data$raceColor + data$dad_education +
                       data$mom_education + data$rural + data$city)
```

Table 15:

	Dependent variable:	
	education	
	(1)	(2)
IV 1	0.242	
	(0.174)	
IV 2		1.006***
		(0.187)
Constant	13.113***	12.529***
	(0.114)	(0.158)
F Statistic	1.926	28.877***
df	1; 998	1; 998
Observations	1,000	1,000
R^2	0.002	0.029
Adjusted R^2	0.001	0.028
Residual Std. Error	2.728	2.690

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

Note that z_1 is not correlated with education, and is thus not an appropriate instrumental variable to estimate the effect of education on wages.

Table 16: Step Two Regression summary

	Dependent variable:
	log(Wages)
Education (z_2)	0.045
	(0.033)
Experience	0.072***
	(0.013)
Expereince^2	−0.003***
	(0.001)
Race (White or Non-white)	−0.245***
	(0.045)
Father's Education	0.006
	(0.006)
Mother's Education	0.022**
	(0.007)
Rural (Yes or No)	−0.096**
	(0.034)
City (Yes or No)	0.191***
	(0.034)
Constant	4.974***
	(0.434)
F Statistic	21.634***
df	8; 714
Observations	723
R^2	0.198
Adjusted R^2	0.189
Residual Std. Error	0.398

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

The results of the regression show that the magnitude of the covariate coefficients is similar using the instrumental variable approach. The coefficient for education has dramatically decreased and is no longer statistically significant.

Question 5: CLM 2

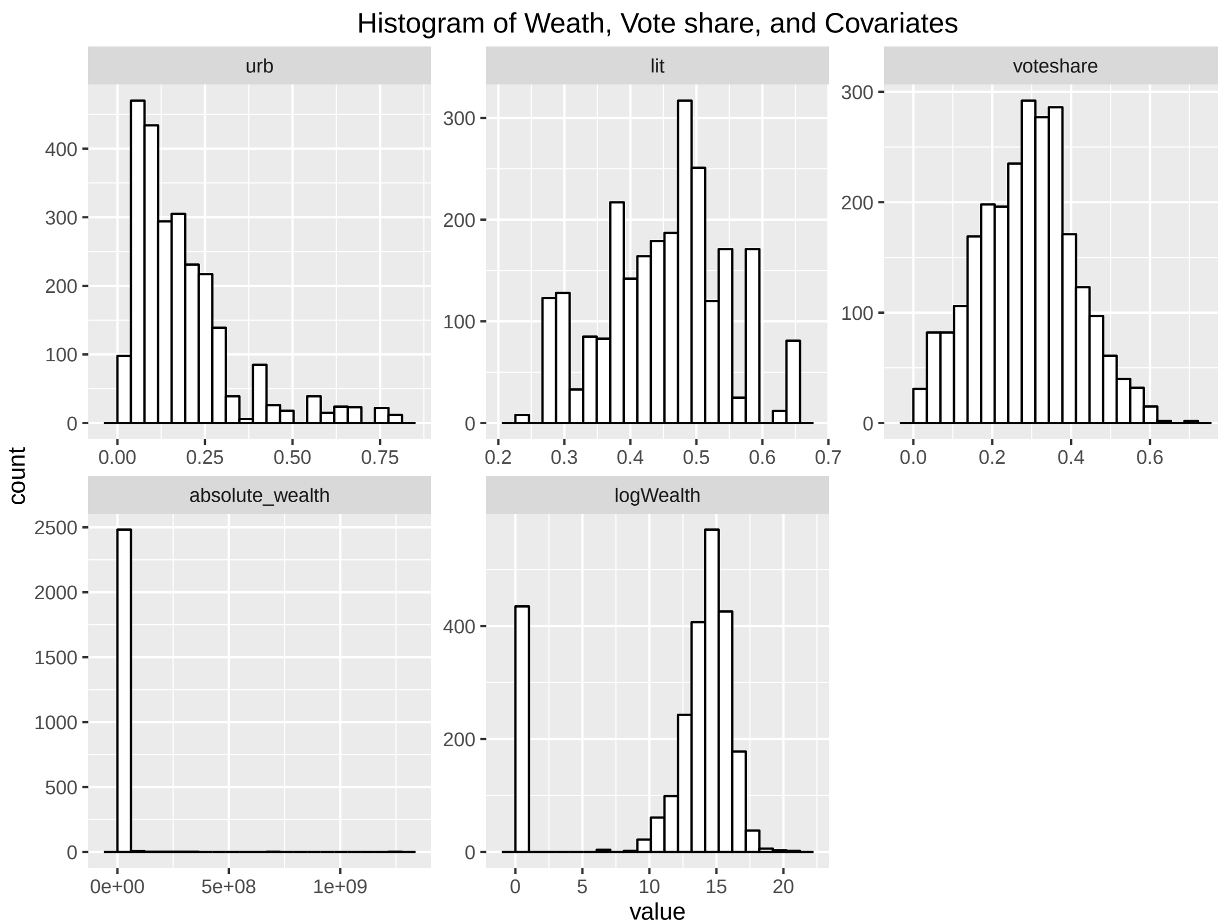
The dataset, `wealthy candidates.csv`, contains candidate level electoral data from a developing country. Politically, each region (which is a subset of the country) is divided in to smaller electoral districts where the candidate with the most votes wins the seat. This dataset has data on the financial 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?

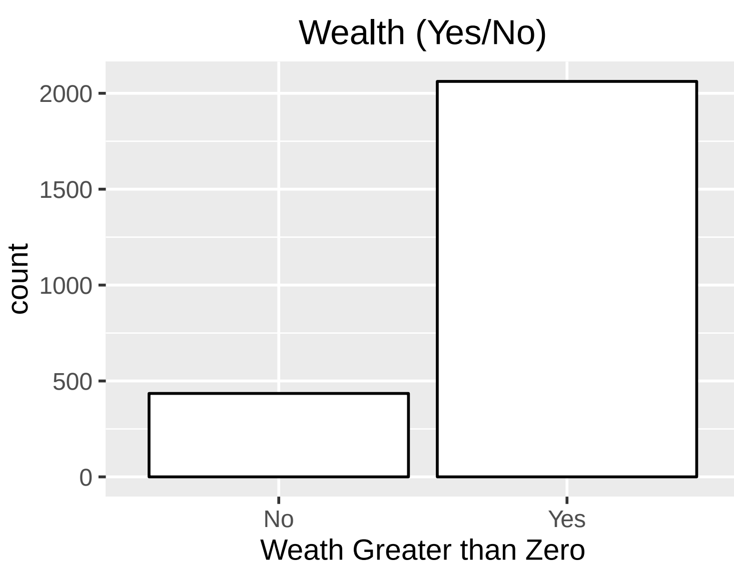
1. **Begin with a parsimonious, yet appropriate, specification. Why did you choose this model? Are your results statistically significant? Based on these results, how would you answer the research question? Is there a linear relationship between wealth and electoral performance?**

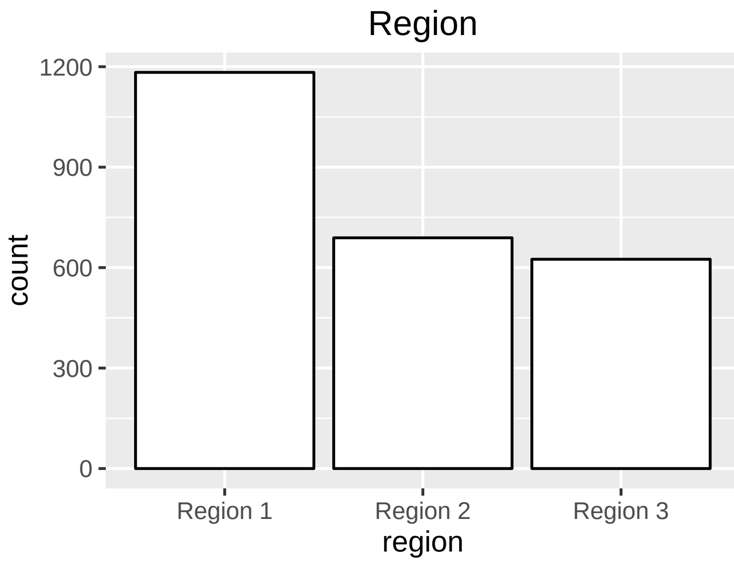
In order to decide on a parsimonious model and dsicover any potential issues with data quality or variable distributions, we will first explore summary statistics and conduct a univariate analysis.

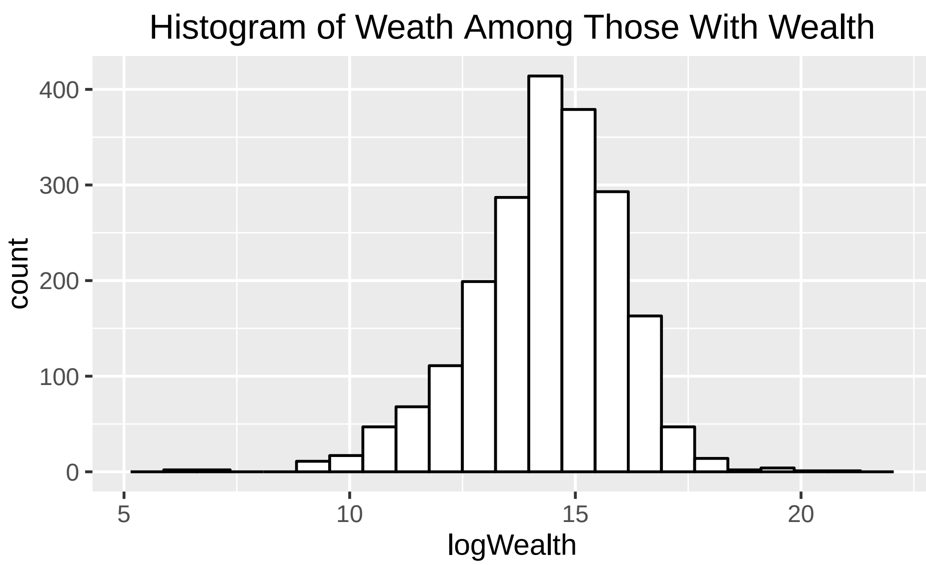
Table 17:

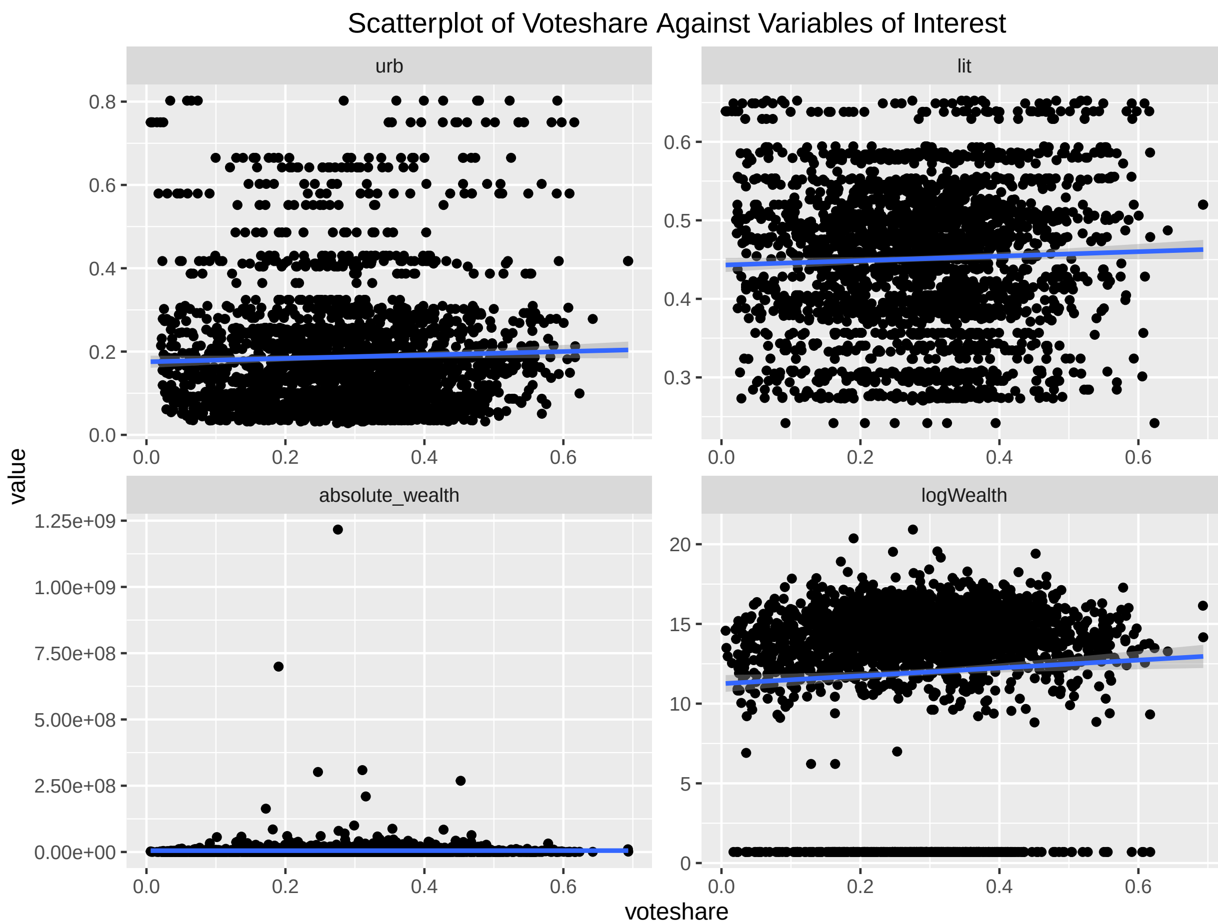
Statistic	X	urb	lit	voteshare	absolute_wealth	logWealth
N	2,497	2,497	2,497	2,497	2,497	2,497
Mean	1,249.930	0.187	0.451	0.288	5,034,105.000	11.961
St. Dev.	721.080	0.149	0.092	0.123	31,098,493.000	5.388
Min	1	0.028	0.242	0.006	2.000	0.693
Pctl(25)	626	0.084	0.385	0.200	187,500.000	12.142
Median	1,250	0.147	0.460	0.293	1,336,629.000	14.106
Pctl(75)	1,874	0.243	0.510	0.368	4,092,001.000	15.225
Max	2,498	0.802	0.652	0.693	1,216,399,232.000	20.919


Histogram of Weath, Vote share, and Covariates
urb
lit
voteshare
absolute_wealth
logWealth
count
value


Wealth (Yes/No)
count
No
Yes
Weath Greater than Zero


Region
1200
900
600
300
0
count
Region 1
Region 2
Region 3
region


Histogram of Weath Among Those With Wealth
400
300
200
100
0
count
5
10
15
20
logWealth



Examining the absolute_wealth and log(absolute_wealth) variables reveals that a large number of observations have the same value. It seems likely that this value codes for having zero absolute wealth. Because we interested in the effect of wealth on voteshare, it seems reasonable to look at this effect among those with wealth greater than zero. By creating a factor variable representing wealth greater than zero, the equation for our parsimonious model will be

$$voteshare = \beta_0 + \beta_1 log(wealth) + \beta_2 hasWealth$$

Table 18: Regression summary

	Dependent variable:
	Voteshare
log(Wealth)	0.005^{**}
	(0.002)
Has Wealth = Yes	-0.057^{*}
	(0.026)
Constant	0.273^{***}
	(0.006)
F Statistic	7.189^{***}
df	2; 2494
Observations	2,497
R^2	0.006
Adjusted R^2	0.005
Residual Std. Error	0.123

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

The parsimonious model reveals a very weak positive effect of wealth on voteshare. The effect is not of practical significance and suggests that there may not be a linear relationship between wealth and voteshare.

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?**

Theoretically, the addition of a quadratic seems questionable because it suggests there is a point at which the returns to wealth stop increasing. Even if the quadratic model resulted in an improvement in the R^2 value, it may be difficult to justify the inclusion of a quadratic term without a theoretical reason.

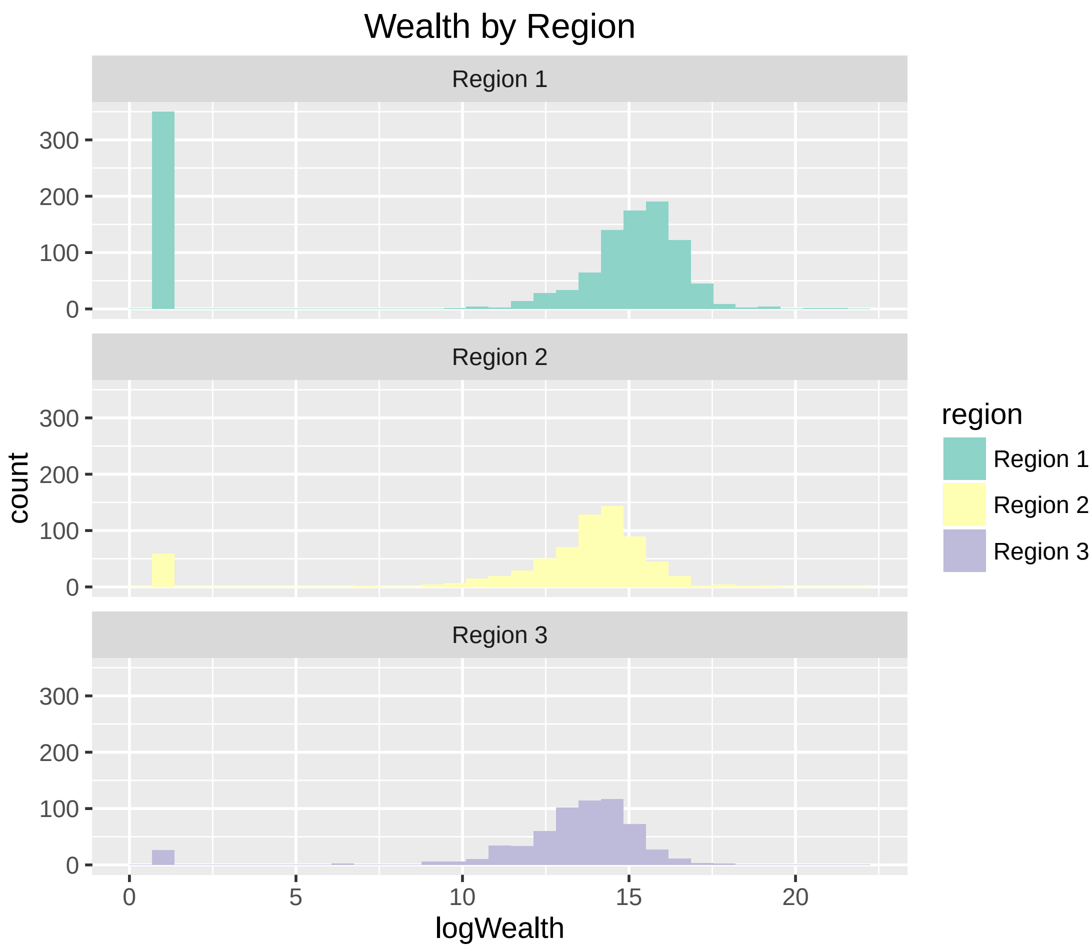
Table 19: Regression summary

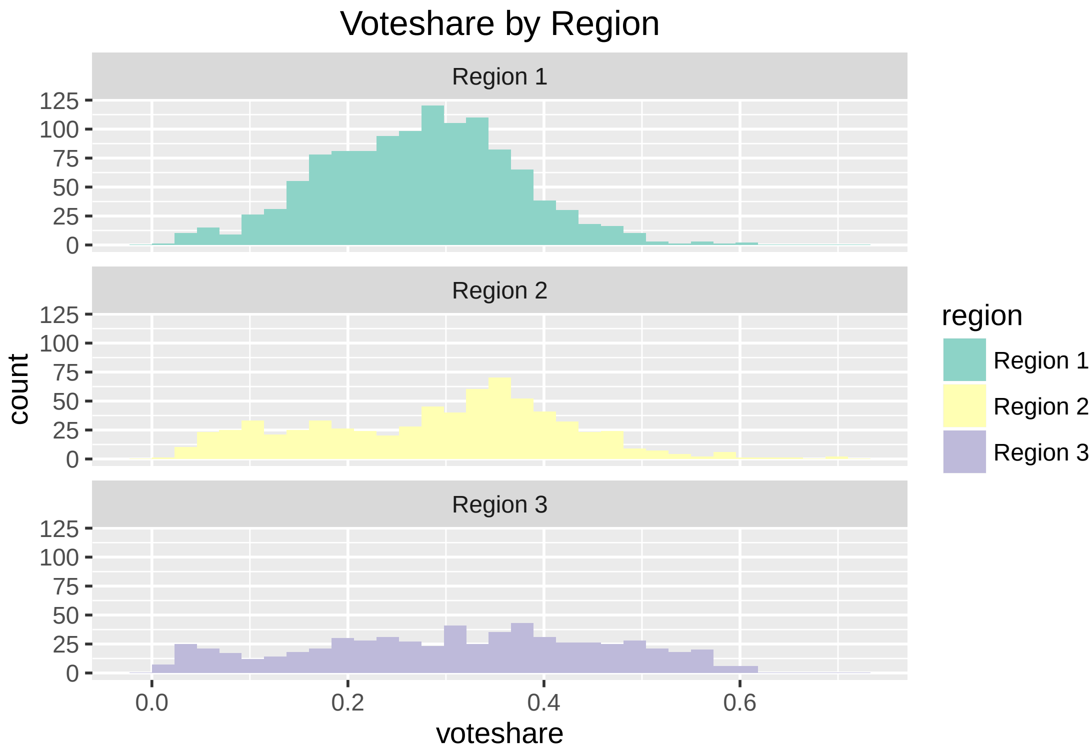
	Dependent variable:
	Voteshare
log(Wealth)	0.005^{**}
	(0.002)
$\log(\text{Wealth})^2$	-0.057^{*}
	(0.026)
Has Wealth = Yes	0.273^{***}
	(0.006)
F Statistic	7.189^{***}
df	2; 2494
Observations	2,497
R^2	0.006
Adjusted R^2	0.005
Residual Std. Error	0.123

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

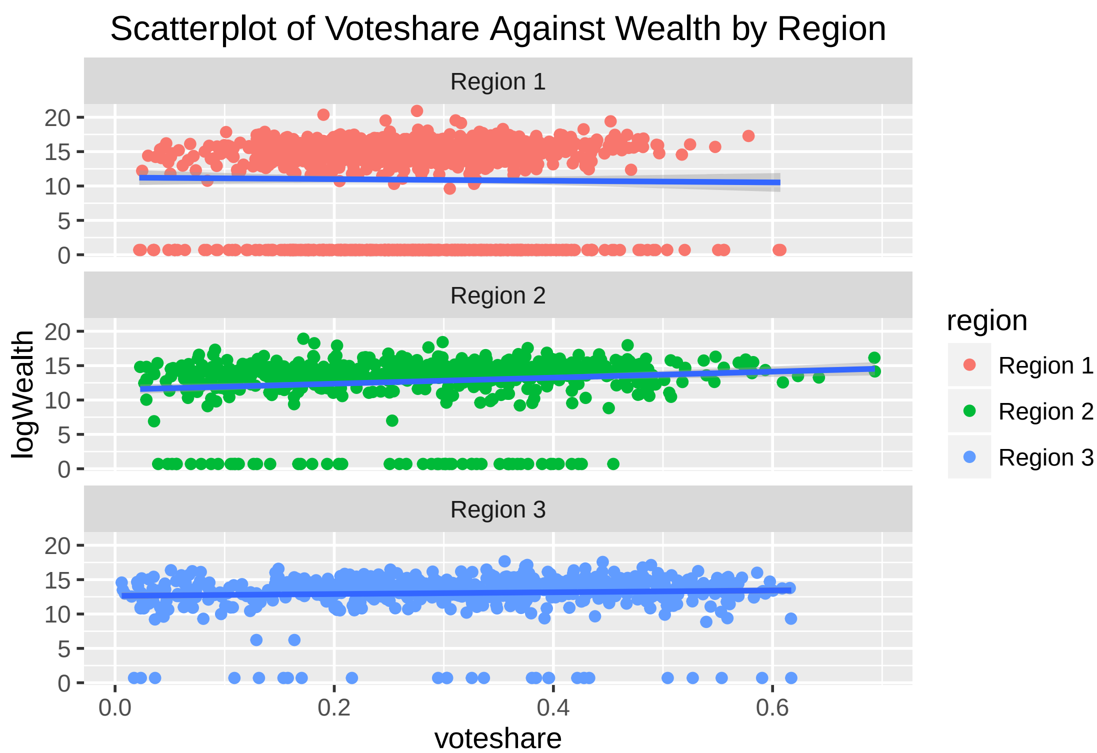
3. **Another team member suggests that it is important to take into account the fact that different regions have different electoral contexts. In particular, the relationship between candidate wealth and electoral performance might be different across states. Modify your model and report your results.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

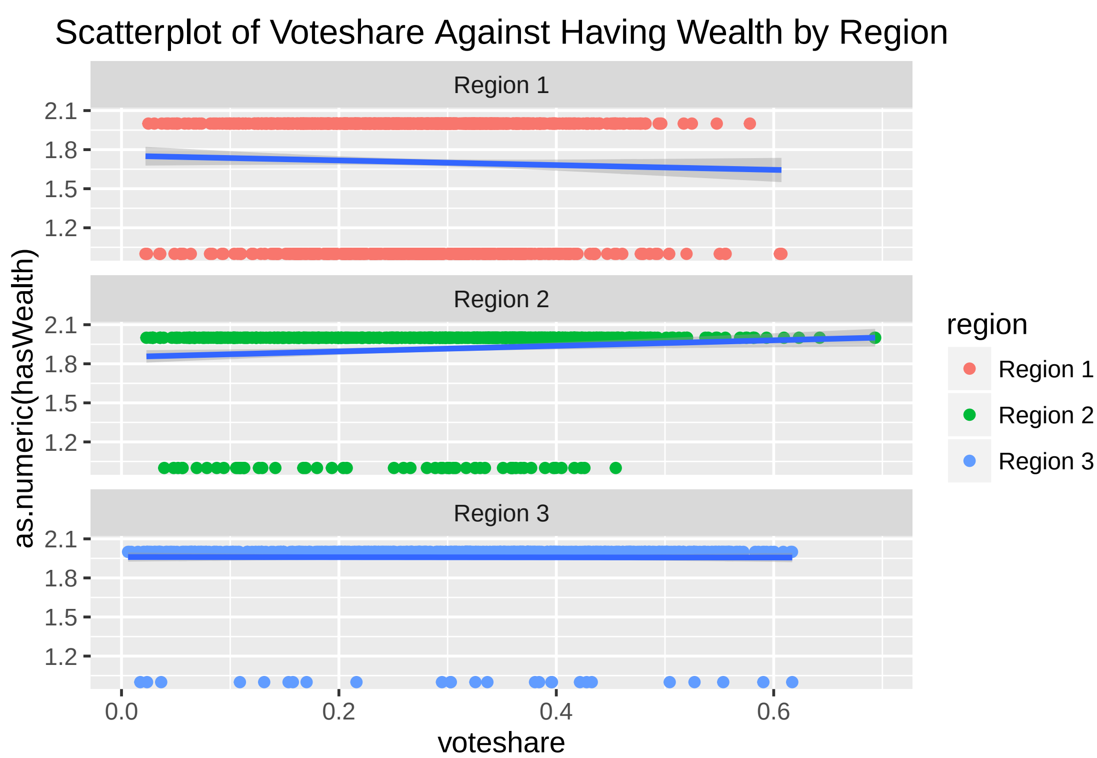
First, we will look at how the variables of interest are distributed across regions.





The histograms of wealth by region and voteshare by region do seem to reveal some geniune structural differences between the regions. The first region has much greater wealth inequality, with a large share of the observations having no wealth and greater average wealth among the wealthy than regions 2 and 3. The voteshare histogram for region one also suggests a difference in election structure, as voteshare percentages tend to be clusered between 15% and 40%, while regions two and three have relatively uniform voteshare distributions between 5% and 60%.





The scatterplots of voteshare against the wealth variables reveal that is all regions, there does not seem to be a strong linear relationship between wealth amd voteshare. There seems to be a level of $log(wealth)$ around 10, below which there are few observations, but once that floor is cleared, there doesn't seem to be a strong relationship between additional wealth and additional voteshare.

Table 20: Regression summary

	Dependent variable:
	Voteshare
log(Wealth)	0.011***
	(0.002)
Has Wealth = Yes	−0.156***
	(0.029)
Region = 2	0.031***
	(0.006)
Region = 3	0.055***
	(0.007)
Constant	0.262***
	(0.006)
F Statistic	19.367***
df	4; 2492
Observations	2,497
R^2	0.032
Adjusted R^2	0.030
Residual Std. Error	0.121

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 21: Model Comparison

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Parsimonious Model	2,494	37.736				
Regions Model	2,492	36.741	2	0.995	33.755	0

The regression output shows that the model including regions accounts for a greater share of the overall variation than the parsimonious model, with the difference between the two models being practically and statistically significant ($F(2, 2492) = 33.755$, $p < 0.001$). This supports the notion that the structural differences in voteshare and wealth between the regions are important to take into account when estimating the effect of wealth on voteshare.

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.**
