Lab 2 Wealthy Candidates

Devesh Tiwari

March 5, 2016

Question 5. Part 1

Begin with a parismonious, 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? In there a linear relationship between wealth and electoral performance?

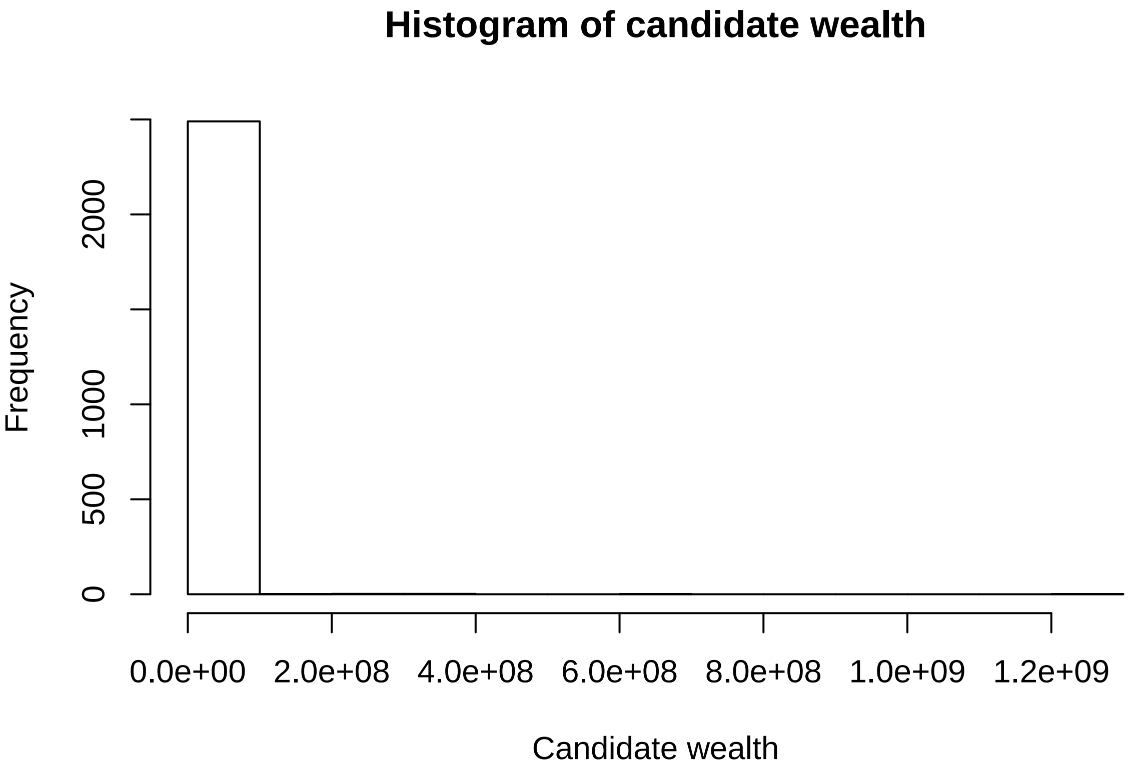
Answer: In this initial model, I regress candidate voteshare on the natural log of candidate wealth.

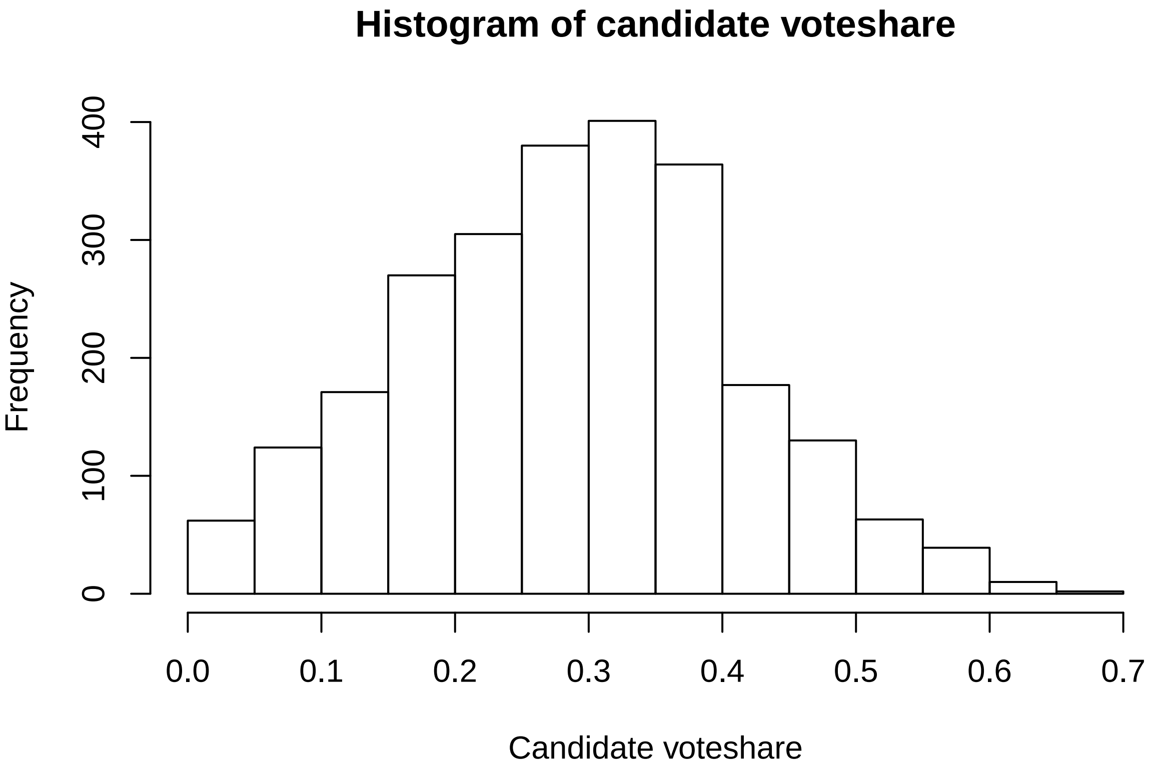
$$voteshare_i = \beta_0 + \beta_1 logWealth_i + \epsilon_i$$

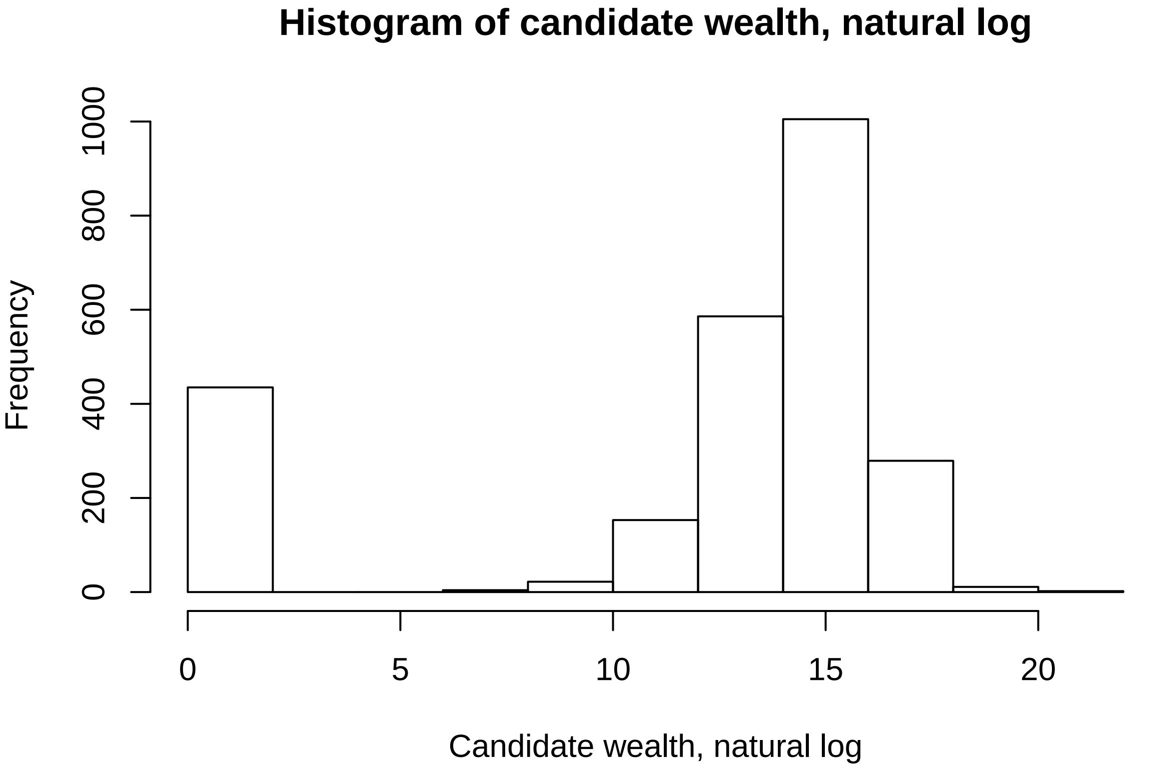
As seen in the exploratory data analysis below, I take the natural log of candidate wealth because wealth is not normally distributed (has extreme outliers). After taking the natural log, I notice that there are a few observations that have a value of 0.69. I am going to assume that these observations represent true errors in the data collection and generation process and I will thus exclude them. Note: If you wanted to include other regressors in this model, you are free to, but you must justify your choice and perform some EDA.

I conduct an OLS analysis using robust standard errors. I find a positive and statistically significant result between candidate wealth and candidate voteshare. A one-unit increase in log wealth corresponds to a 0.5 percentage point increase in candidate voteshare. Given that the standard deviation of log wealth is 1.6, these results suggest that it is possible that some candidates are wealthy enough to gain neccessary voteshare and win very tight elections. Having said that, I do not have enough information to assess whether candidates with unequal levels of wealth faced each other in tight elections. Futhermore, as a later discussion illustrates, I do not believe that these results are unbiased or causal.

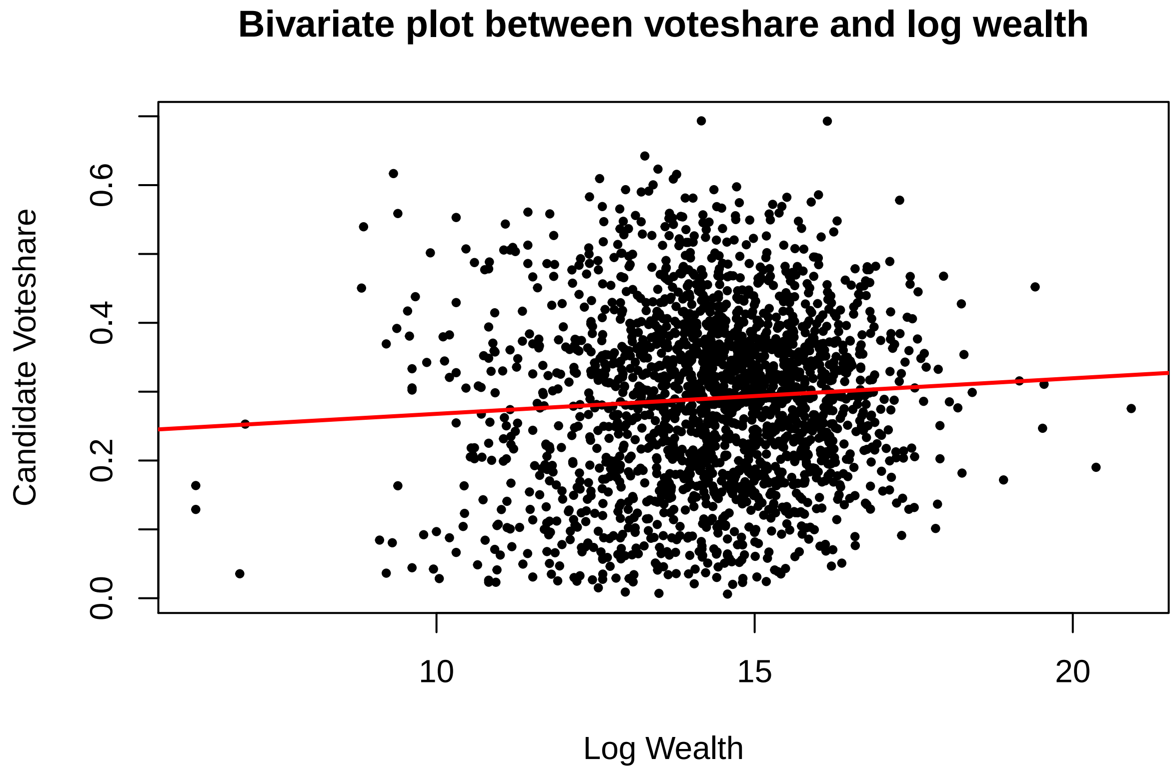
The relationship between wealth and voteshare is not linear, it is log-linear.







```
##
## ============================================================================
## Statistic          N       Mean         St. Dev.      Min          Max
## ----------------------------------------------------------------------------
## urb              2,062     0.182          0.149       0.028          0.802
## lit             2,062     0.456         0.094       0.242        0.652
## voteshare       2,062     0.290         0.125       0.006        0.693
## absolute_wealth 2,062 6,096,101.000 34,128,597.000 501.000 1,216,399,232.000
## lw              2,062    14.338         1.646       6.217       20.919
## ----------------------------------------------------------------------------
```



```
##
## Relationship between candidate wealth and electoral performance. Ordinary Least Squares
## ====================================
##               Dependent variable:
##           ---------------------------
##
## ------------------------------------
## lw                0.005***
##                    (0.002)
##
## Constant          0.216***
##                    (0.027)
##
## ====================================
## ====================================
## Note:    *p<0.1; **p<0.05; ***p<0.01
```

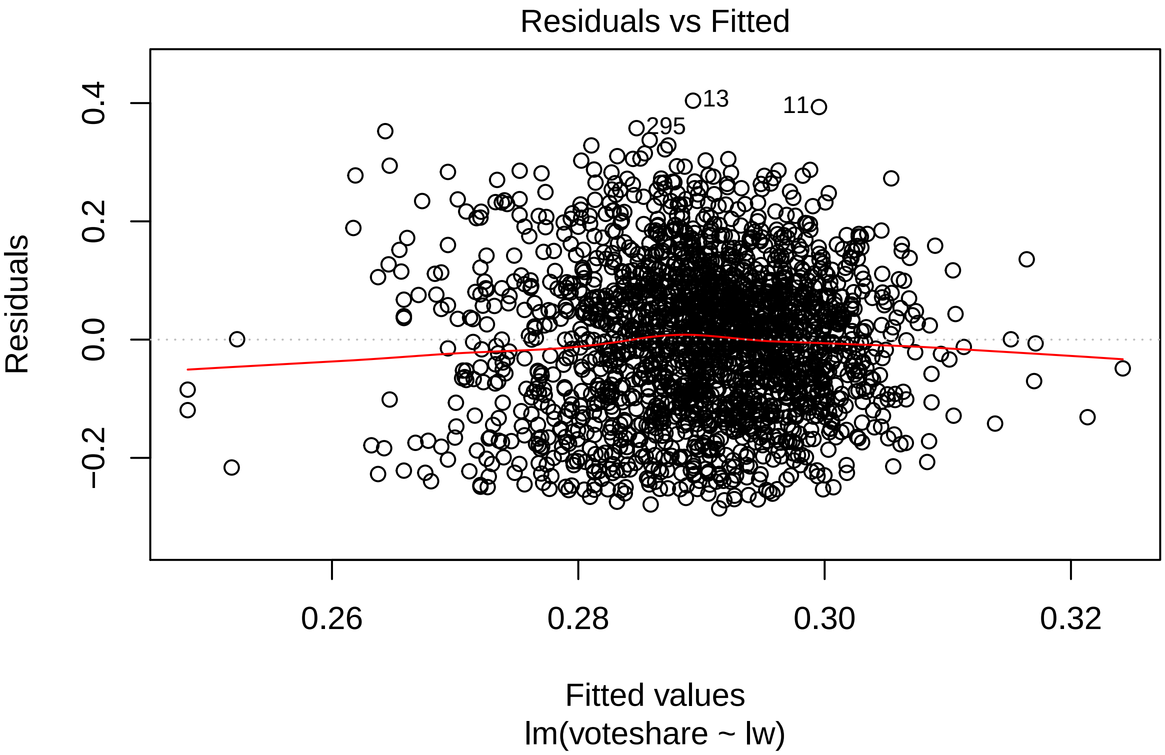
Question 5: Part 2.

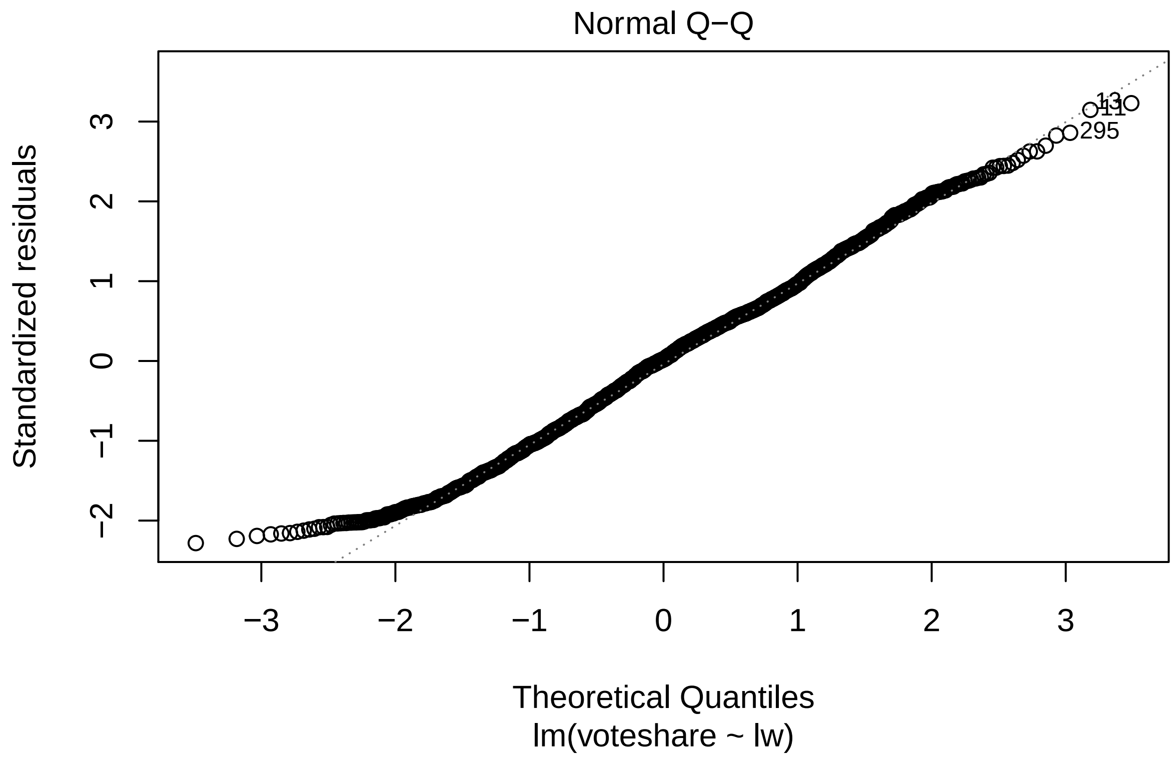
A team member suggests adding a quadratic term to your regression. Based on your prior model, is such an addition warranted? Add this term and interpret your results. Do wealthier candidates fare better in elections?

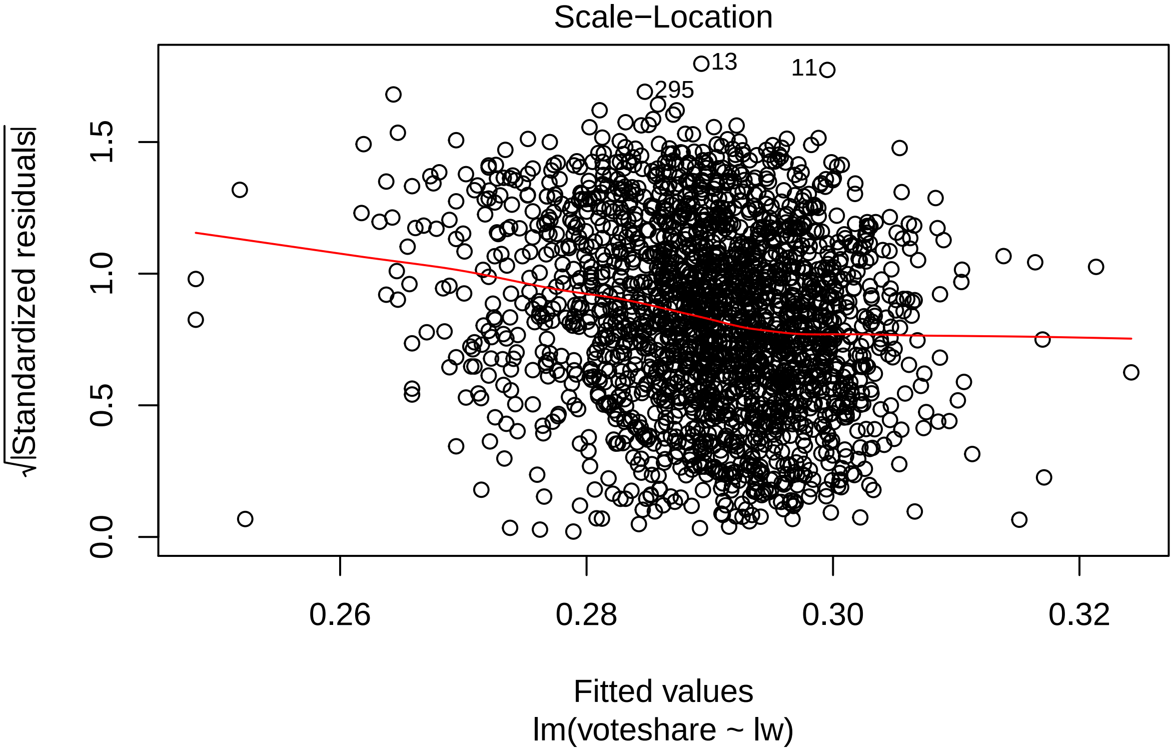
Answer: After examining the residual plots from my initial model, it seems as if the residuals are normally distributed, though there is some visual evidence suggesting that the zero conditional mean assumption has been violated. It is possible that this model is mis-speccificed and thus the inclusion of a quadratic term is worth exploring. As a matter of logic, one might also believe that there are diminshing returns to the use of wealth in elections: After a certain level of wealth, its marginal impact might become smaller.

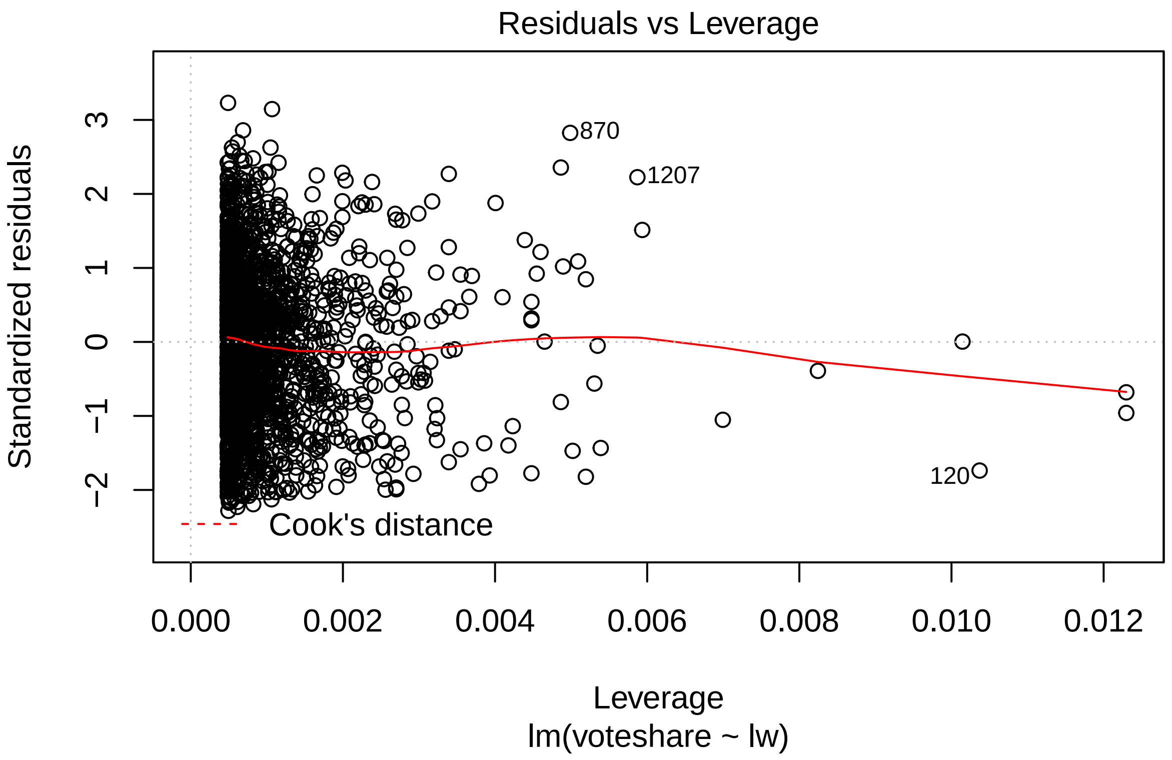
The OLS model shows some evidence that there are diminsihing returns to wealth as the coefficient on log wealth is positive and the coefficient on the quadratic term is negative (both are statistically significant at the 0.05 level). Given that the only regressors are wealth and its square, I need to see some predicted values before I can fully comment on the impact of wealth on voteshare. According to this plot, candidate voteshare seems to peak when candidates have a logged wealth value of 15 and voteshare decreases from there. In order to see if this is really the case, I produce a scatterplot of wealth and voteshare, and then overlay the predicted value of voteshare from both the quadratic model and linear model. I am skeptical that after a certain threshold, wealth negatively impacts voteshare and conclude (for now) that this is an artificat of using a quadtratic term.

The residual plots for the quadratic model shows that the zero condional mean assumption seems to be still be violated and that residual v fitted values plot is clustered on the right hand side. In other words, because voteshare actually has a maximum value near 0.30, there are no fitted values greater than that. Was adding a quadratic term worth including? On the one hand, both the linear and quadtradic models do not seem to fit the data very well, while on the other hand, an F-test shows that the inclusion of the quadratic term does increase the explanaotry power of the model (though not by very much.)

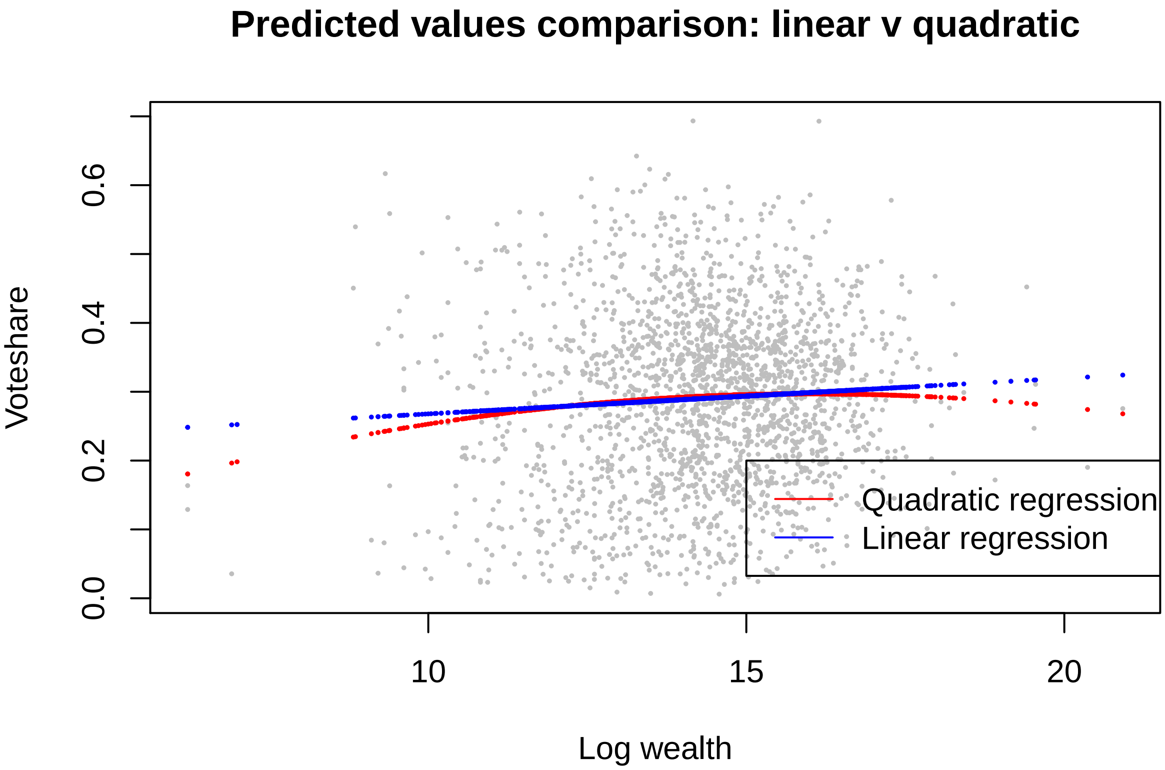


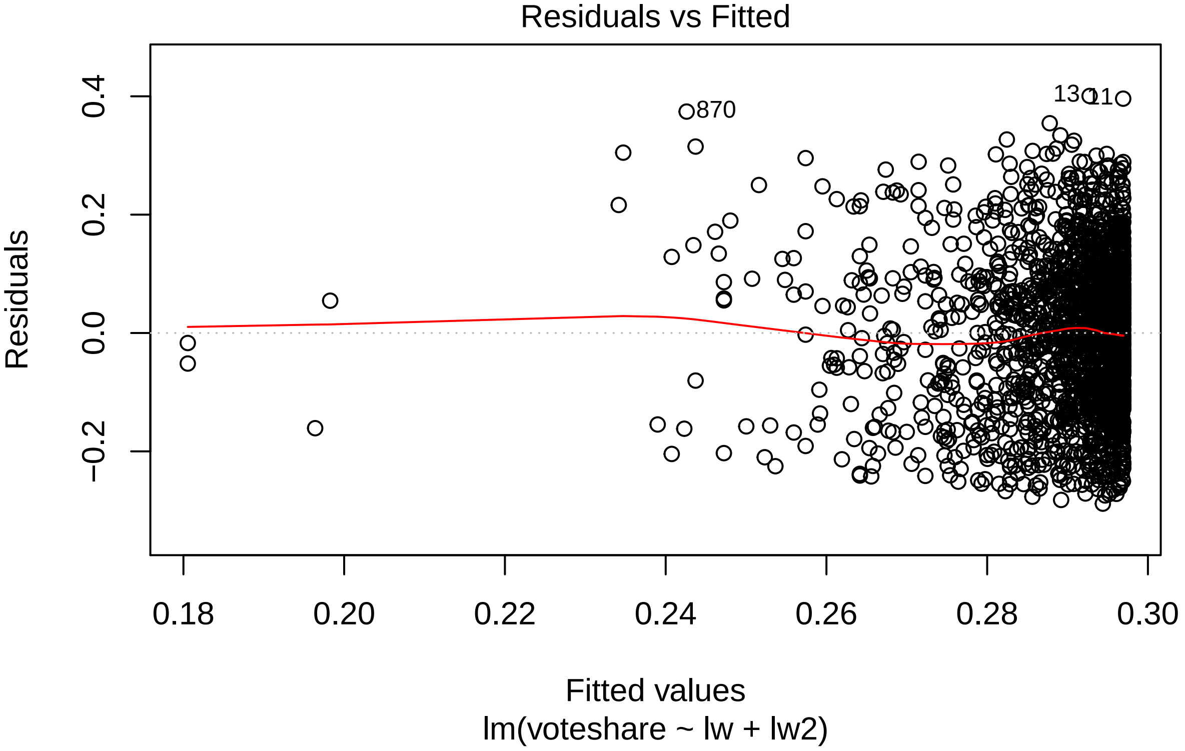

Normal Q−Q
Standardized residuals
−2
−1
0
1
2
3
−3
−2
−1
0
1
2
3
13
11
295
Theoretical Quantiles
lm(voteshare ~ lw)

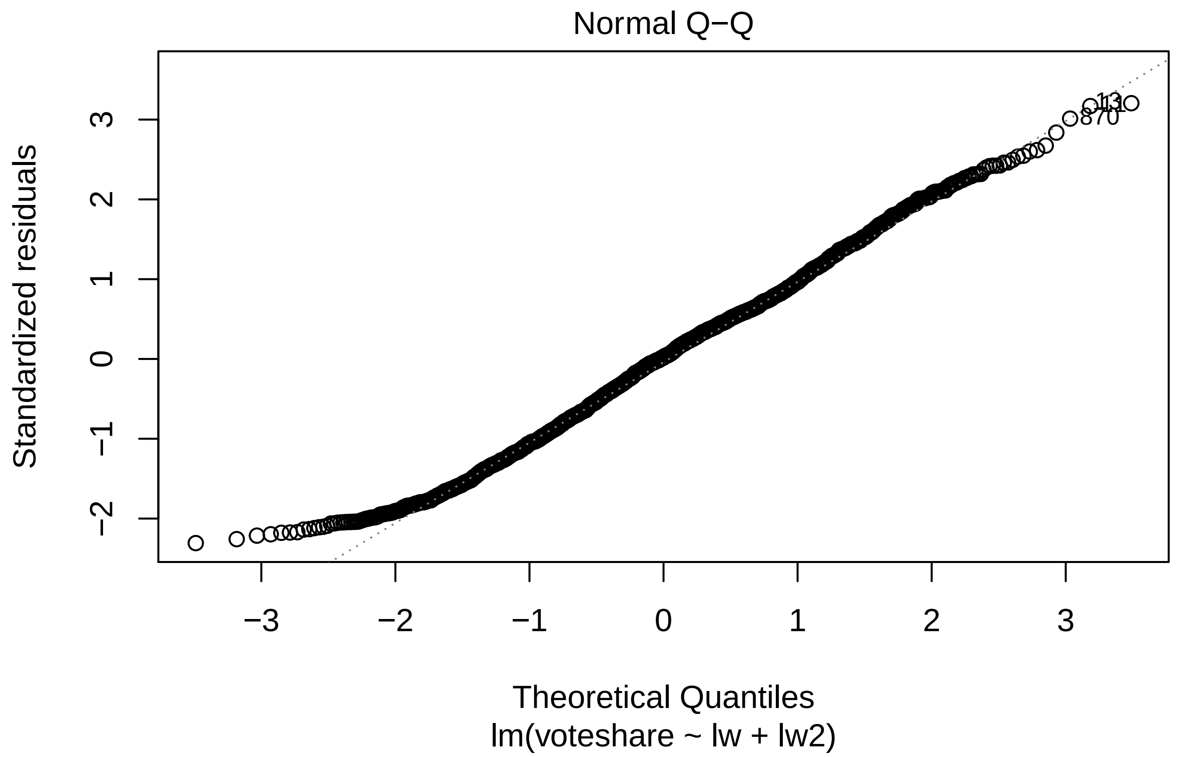

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.26
0.28
0.30
0.32
13
11
295
Fitted values
lm(voteshare ~ lw)

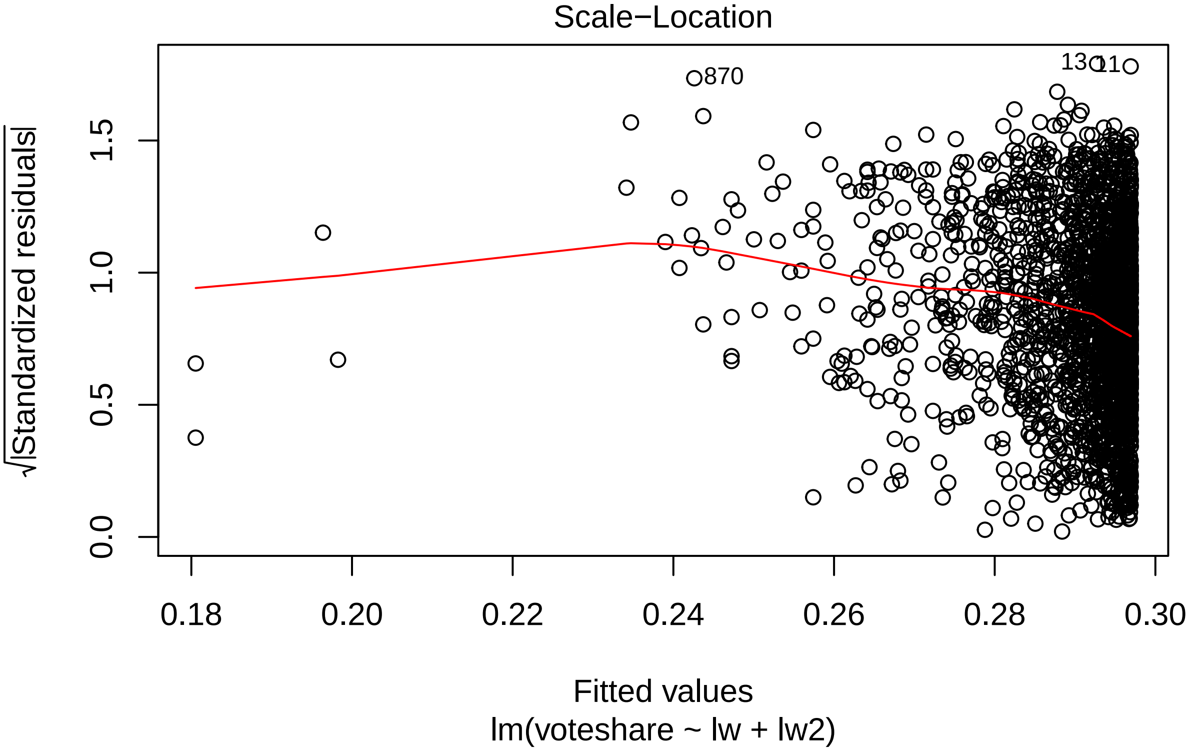


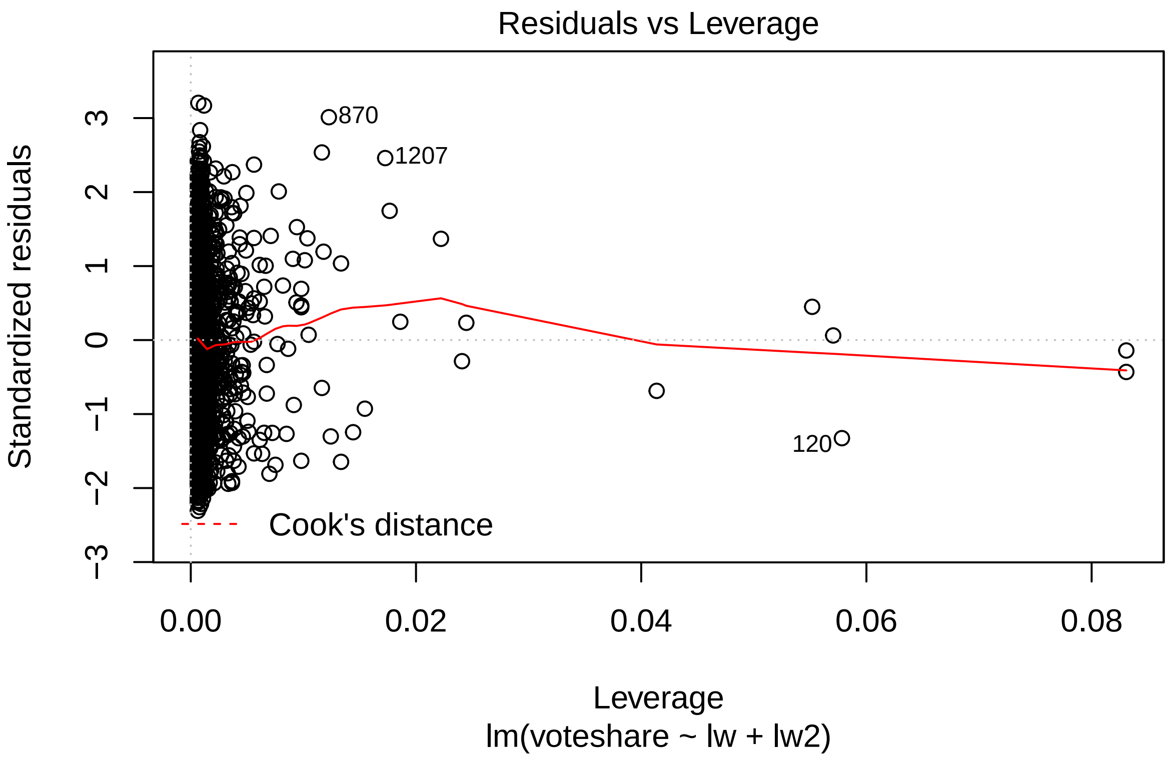
```
##
## Candidate wealth and electoral performance, quadratic
## ====================================
##                Dependent variable:
##           ---------------------------
##
## ------------------------------------
## lw                    0.039**
##                       (0.017)
##
## lw2                  -0.001**
##                       (0.001)
##
## Constant              -0.014
##                       (0.122)
##
## ====================================
## ====================================
## Note:    *p<0.1; **p<0.05; ***p<0.01
```


Predicted values comparison: linear v quadratic
Voteshare
0.0
0.2
0.4
0.6
10
15
20
Log wealth
Quadratic regression
Linear regression


Residuals vs Fitted
Residuals
−0.2
0.0
0.2
0.4
0.18
0.20
0.22
0.24
0.26
0.28
0.30
Fitted values
lm(voteshare ~ lw + lw2)
870
130
11


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
13
11
870
Theoretical Quantiles
lm(voteshare ~ lw + lw2)


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
870
13
11
0.18
0.20
0.22
0.24
0.26
0.28
0.30
Fitted values
lm(voteshare ~ lw + lw2)



```
## Wald test
##
## Model 1: voteshare ~ lw
## Model 2: voteshare ~ lw + lw2
##   Res.Df Df      F  Pr(>F)
## 1   2060
## 2   2059  1 4.1411 0.04198 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 5. Part 3

Another team member suggests that it is important to take into account the fact that different regions have different electoral contexts. In par- ticular, 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.

Answer: Your EDA and residual analyses should have revealed the importance of studying regional impacts. In order to fully explore whether the relationship between candidate wealth and voteshare is different across regions, you need to add an indicator variable for two regions and interact them with wealth (and wealth squared if you feel that it is an important thing to consider.) In order to fully evaluate the impact of region, the following models need to be run:

(1) Base model
(2) Quadratic model
(3) Base model + Regional Indicators
(4) Quadratic model + Regional Indicators
(5) Base model + Regional Indicators + Interaction

(6) Quadratic model + Regional Indiacors + Linear Interactions
(7) Quadratic model + Reginal Indicators + Linear Interaction + Quadratic Interaction.

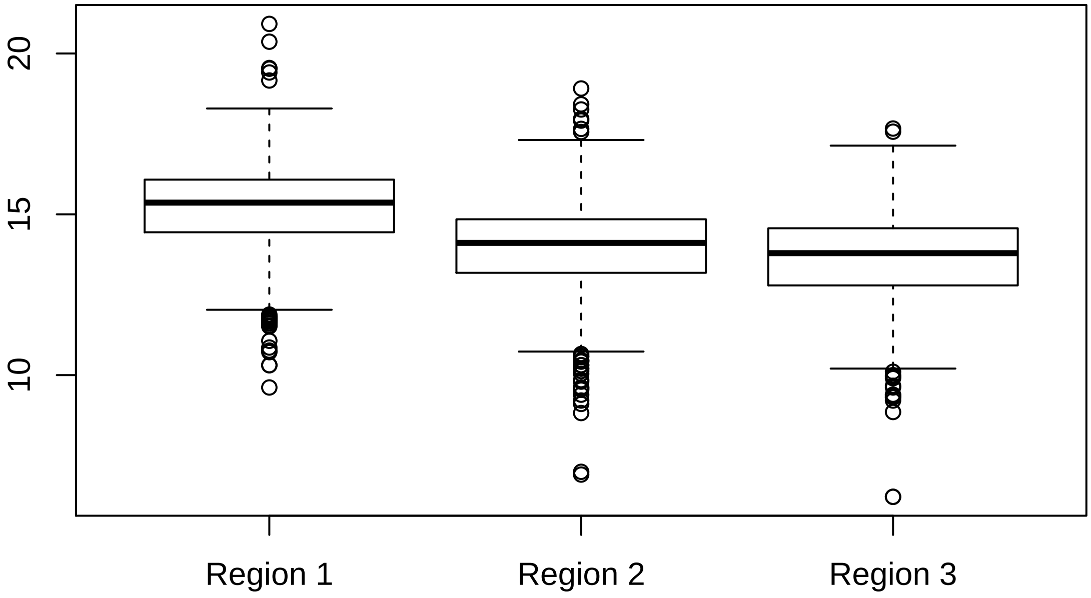
In order to fully answer the question, we need to test whether any of the interaction terms are neccessary. Before we even get to this step, we need to determine whether or not the regional dummy variables are neccessary. In order to evaluate that, we need to conduct a series of F-tests.

If you believe that the addition of the quadrtic term is/was neccesary, then you need to conduct an F-test comparing Model (4) with Model (2); if you feel that the quadratid was not neccessary, then you need to compare Model (3) with Model (1).

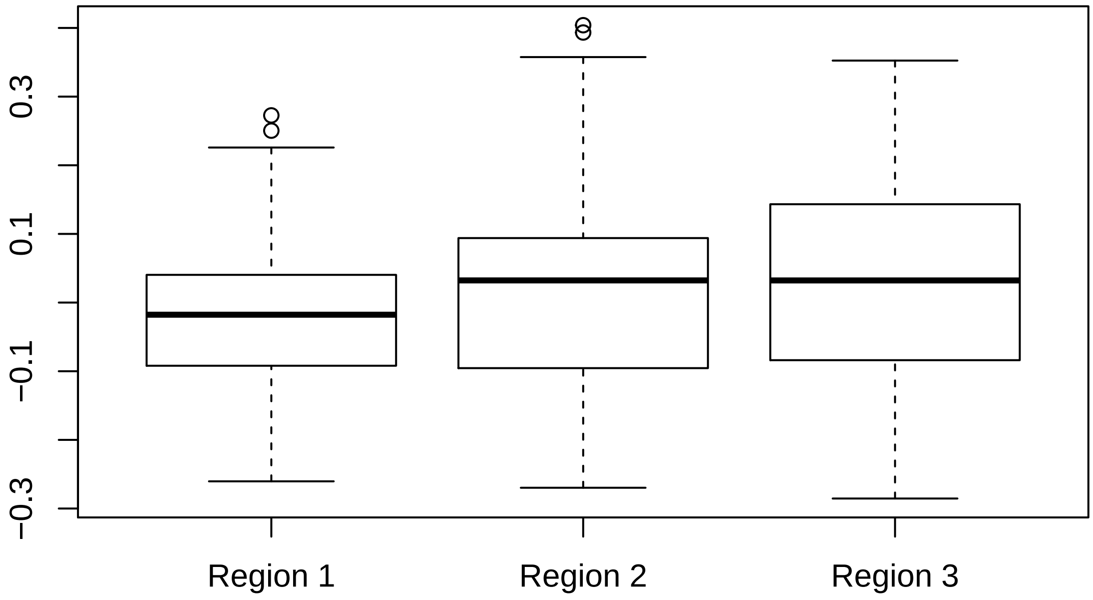
The first F-test compares Model (4) with Model (2) and it passes.The addition of the regional dummy variables increases the explanatory power of the model. When regional indicators are used, the statistical significance of wealth and its quadratic term vanishes, though wealth is still statistically significant when the quadratic term is not included. As we can see here, our ability to answer the research question really does hinge on how we specify our model. If we thnk the inclusion of a quadratic term is justified, then we would conclude that wealth and electoral performance are unrelated to one another and that the relationship is driven by regional level differences. Note the temptation to go back and say, "Just kidding! I really don't believe that we need to include a quadratic term!"

Moving on, recall that we have to examine whether the relationship between wealth and electoral performance differes by region, and in order to do that we need to include interaction terms. Our next F-test compares Model (6) with Model (4), and if that passes we compare Model (7) with Model (6). According to the results below, the inclusion of linear interaction terms with region does not increase the model's explanatory power. As a result, we would fail to reject the hypothesis that wealth and voteshare have a different relationship across regions.

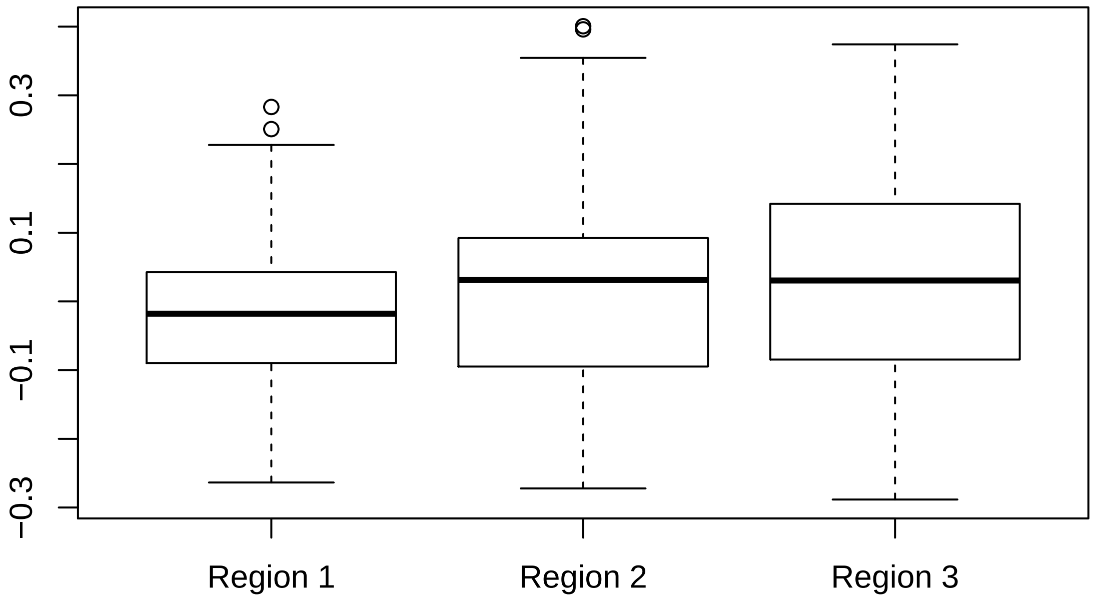
```
# Are regional effects needed? EDA and residual analysis
boxplot(wc$lw~wc$region)
```



```
# Residuals
boxplot(baseModel$residuals~wc$region[!is.na(wc$region)])
```



```
boxplot(quadraticModel$residuals~wc$region[!is.na(wc$region)])
```



```
### Build models
baseRegionModel<- lm(voteshare~lw+region, data = wc) #3
quadraticRegionModel<- lm(voteshare~lw+lw2+region, data=wc) #4
baseRegionInteractionModel<- lm(voteshare~lw + region + lw:region, data=wc) #5
quadraticRegionLinearInteractionModel<- lm(voteshare~lw + lw2 + region + lw:region, data=wc) #6
quadraticRegionQuadInteractionModel<-
lm(voteshare~lw+lw2+region+lw:region+lw2:region, data=wc) #7





#Robust Standard Errors
```

```r
baseRobust<-coeftest(baseModel, vcov=vcovHC)

quadRobust<-coeftest(quadraticModel, vcov=vcovHC)

baseRegionRobust<-coeftest(baseRegionModel, vcov=vcovHC)

quadRegionRobust<-coeftest(quadraticRegionModel, vcov=vcovHC)

baseRegionInteractRobust<-coeftest(baseRegionInteractionModel, vcov=vcovHC)

quadRegionLinearInteractRobust<- coeftest(quadraticRegionLinearInteractionModel, vcov=vcovHC)

quadRegionQuadInteractRobust<- coeftest(quadraticRegionQuadInteractionModel, vcov=vcovHC)


stargazer(baseRobust, quadRobust, baseRegionRobust, quadRegionRobust, type="text")
```

```
## 
## ==================================================
##                         Dependent variable:       
##                 ----------------------------------
##                                                   
##                   (1)      (2)      (3)      (4)  
## --------------------------------------------------
## lw              0.005*** 0.039**  0.012***  0.027 
##                 (0.002)  (0.017)  (0.002)  (0.017)
##                                                   
## lw2                      -0.001**           -0.001
##                          (0.001)           (0.001)
##                                                   
## regionRegion 2                    0.041*** 0.040***
##                                   (0.007)  (0.006)
##                                                   
## regionRegion 3                    0.061*** 0.060***
##                                   (0.007)  (0.007)
##                                                   
## Constant        0.216***  -0.014  0.088***  -0.017
##                 (0.027)  (0.122)  (0.030)  (0.124)
##                                                   
## ==================================================
## ==================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```r
waldtest(quadraticModel,quadraticRegionModel, vcov=vcovHC)
```

```
## Wald test
## 
## Model 1: voteshare ~ lw + lw2
## Model 2: voteshare ~ lw + lw2 + region
##   Res.Df Df      F    Pr(>F)    
## 1   2059                        
## 2   2057  2 39.827 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
stargazer(baseRegionRobust, quadRegionRobust, quadRegionLinearInteractRobust,
          quadRegionQuadInteractRobust, type="text")
```

```
##
## ====================================================
##                           Dependent variable:
##                    ---------------------------------
##
##                      (1)      (2)      (3)     (4)
## ----------------------------------------------------
## lw                 0.012***  0.027    0.023  -0.001
##                    (0.002)  (0.017)  (0.021) (0.030)
##
## lw2                         -0.001  -0.0004 0.0004
##                             (0.001)  (0.001) (0.001)
##
## regionRegion 2     0.041*** 0.040***  0.038  -0.272
##                    (0.007)  (0.006)  (0.069) (0.322)
##
## regionRegion 3     0.061*** 0.060***  0.027  -0.135
##                    (0.007)  (0.007)  (0.077) (0.329)
##
## lw:regionRegion 2                    0.0001   0.044
##                                      (0.005) (0.046)
##
## lw:regionRegion 3                     0.002   0.024
##                                      (0.005) (0.047)
##
## lw2:regionRegion 2                            -0.002
##                                               (0.002)
##
## lw2:regionRegion 3                            -0.001
##                                               (0.002)
##
## Constant           0.088***  -0.017   0.017   0.195
##                    (0.030)  (0.124)  (0.160) (0.222)
##
## ====================================================
## ====================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

```r
waldtest(quadraticRegionModel, quadraticRegionLinearInteractionModel, vcov=vcovHC)
```

```
## Wald test
##
## Model 1: voteshare ~ lw + lw2 + region
## Model 2: voteshare ~ lw + lw2 + region + lw:region
##   Res.Df Df      F Pr(>F)
## 1   2057
## 2   2055  2 0.1098  0.896
```

Question 5. Part 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?

Answer: If candidate wealth is exogenous and if it is possible to manipulate it exogenously, then we would have a causal and unbiased estimate. It is unlikely that candiate wealth is unrelated to any unobserved variables. It might be the case, for example, that political skill is tied to private wealth and thus candidates who do well in elections tend to be wealthier as well.

It might be possible to manipulate candidates' wealth exogenously if the change were to be small (i.e. sending very small amounts of money to candidates increases their wealth, but not by much). However, it seems unlikely that we could exogenously change their wealth levels without impacting other variables. For exapmle, if we were to increase some candidates' wealth by one standard deviation (in logged terms), it seems likely that that would catch the interest of the authorities or other political parties, thus impacting their electoral performance.

Question 5. Part 5

Someone proposes a diference in difference design. Please write the equation for such a model. Under what circumstances would this design yield a causal effect?

The only design that makes sense would involve having similar data tying candidate wealth to perofrmance in a prior election:

$$voteshareChange_i = \alpha_0 + \alpha_1 wealthChange_i + \mu_i$$

This model would yield an unibased and causal estimate if the change in candidate wealth were exogenous and if there were no time-variant error. This seems unlikely on both fronts. First, candidates who won their prior election might end up wealthier than candidates who lost. If this were the case, then the change in candidate wealth is not exogenous. Furthermore, if there is a long gap between the two elections, then the political climate could have changed such that the underlying relationship between wealthy and performance is different as well.