Writing Sample:

The 2016 US presidential election is the most recent and controversial presidential election in recent years. The geographical patterns that appear in the voting behavior allow us to wonder whether there is a geographical voting pattern in the election and whether the difference in demographic and economic characteristics is the cause of the difference in the voting behavior. The purpose of this paper is to investigate the above issues and to determine the role of each population and economic characteristics in influencing the voting behavior. This problem is studied by proposing a series of multivariate regressions, which include variables as a series of demographic and economic characteristics and as a set of non-measurable region-specific characteristics.

In the empirical analysis, this paper first vividly presents the data and the relationship between variables in images. Then we use the correlation analysis to study the relationship between voting behavior and demographic and economic characteristics. And the linear regression model and the mixed effects model are used to establish four models which further analyze the two. After obtaining the results, the paper analyzes the performance of the four models and the facts reflected. The paper build relationship between the county's voter behavior and the county's demographic and economic characteristics, and discover the role of the non-measurable regional characteristics in influencing the voting behavior.

1. Data

We gather poll results data[1] for each counties. The variables are shown in Table 1,

Table 1 Poll Results and Index

Variable Name	Variable Meaning	Variable Type
fips	index for the county	numeric
Clinton_raw	the percentage of voters in the county who vote for Clinton in 2016 US Election	numeric
Trump_raw	the percentage of voters in the county who vote for Trump in 2016 US Election	numeric
total_votes_2016	Total votes in the county in 2016 US Election	numeric
Obama_raw	the percentage of voters in the county who vote for Obama in 2012 US Election	numeric
Romney_raw	the percentage of voters in the county who vote for Romney in 2016 US Election	numeric

We also gather economic and demographic data[2] of each county. We take out the data for the Alaska as it don't contain county level data. The variables are shown in Table 2,

[1] https://github.com/tonmcg/County_Level_Election_Results_12-16

[2] https://github.com/benhamner/2016-us-election/tree/master/input/county_facts_saved

Table 2 Economic and Demographic Data

Variable Name	Variable Meaning	Variable Type
1. fips	index for the county	numeric
2 . population2014	population of the county in 2014	numeric
3 . Edu_batchelors	Percentage of people who get Bachelor's degree in the county	numeric
4 . White	Percentage of white people in the county	numeric
5 . Black	Percentage of black people in the county	numeric
6 . Hispanic	Percentage of Hispanic people in the county	numeric
7 . Income	average income of the county in the last 12 months	numeric
8 . Density	density of the county	numeric
9 . Poverty	Percentage of people below the poverty line in the county	numeric
10. state_abbr	state where the county is located	string
11. Region	the region the county is located, including South, West, Northeast, and Midwest	string
12. Sub.Region	the sub-region the county is located, including East North Central, East South Central, Middle Atlantic ,Mountain, New England, Pacific, South Atlantic, West North Central, and West South Central	string

2. Descriptive Statistics and Data Visualization

We cast the results of the polls of each county to the map. See figure 1,

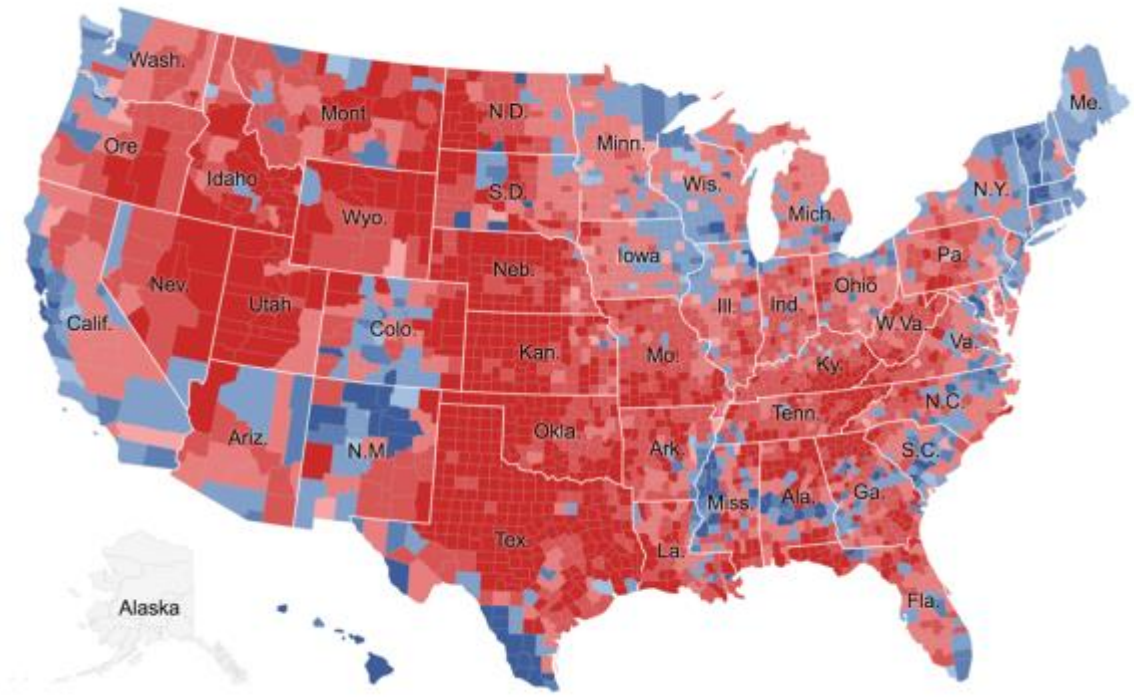

Figure 1 Poll results of each county in 2016 US election

The deeper the blue for the county, the more dominant the Democratic is in the county. And the deeper the red for the county, the more dominant the Republican is in the county.

This map to some extent explains why Trump can win the 2016 US Election. Although Trump win fewer votes than Clinton, votes for Clinton is concentrated in a few states, whereas votes for Trump is spread across large geographic areas, which led to his Electoral College winning.

From the figure, we can roughly see that the votes for Clinton come from the more developed and densely populated coastal areas, while votes for Trump come from central and western areas of relatively sparse population.

The descriptive statistics for Clinton_raw and Trump_raw is shown in Table 3.

Table 3 Descriptive statistics for Clinton_raw and Trump_raw

Variable	Mean	Std	Q1	Q3	skewness
Clinton_raw	0.3170703	0.1535777	0.2047592	0.3996103	0.950352
Trump_raw	0.6361523	0.1564989	0.5494785	0.7514706	-0.8369809

The histogram for Clinton_raw and Trump_raw is shown in Figure 2 and Figure 3

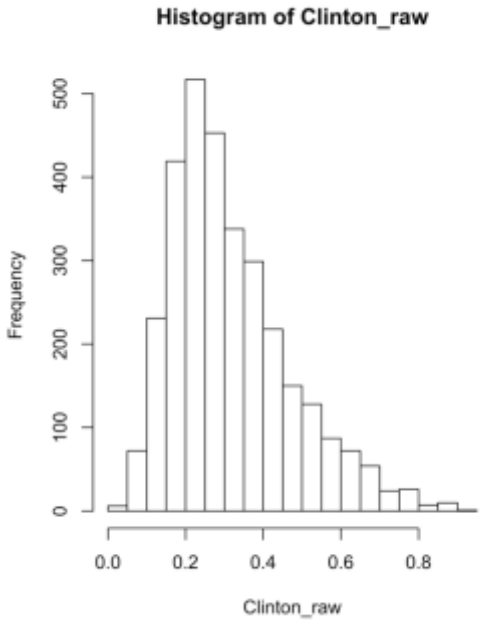

Figure 2 histogram for Clinton_raw

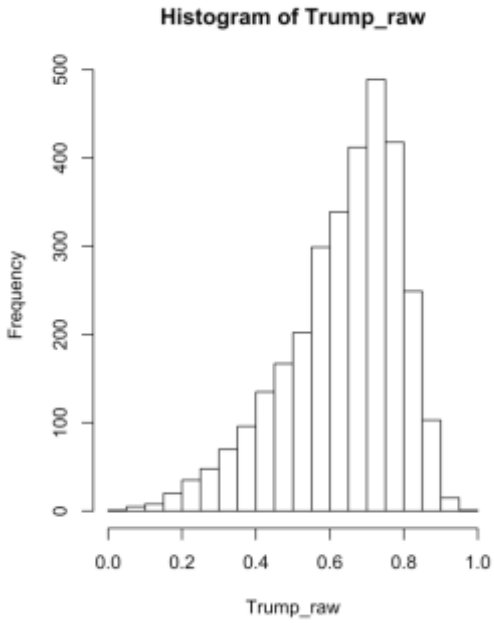

Figure 3 histogram for Trump_raw

From the histogram we can see that both Clinton_raw and Trump_raw is skewed. Clinton_raw is right skewed and Trump_raw is left skewed. The skewness of Clinton_raw is 0.950352, and the skewness of Trump_raw is -0.8369809.

We transform the data using the following expression to reduce the skewness. After the transformation, the data should approximately fit Gaussian distribution.

$$\text{Clinton} = -\log\frac{1-\text{Clinton}\_raw}{\text{Clinton}\_raw}$$

$$Trump = -\log\frac{1-Trump\_raw}{Trump\_raw}$$

The descriptive statistics for Clinton and Trump is shown in Table 4.

Table 4 Descriptive statistics for Clinton and Trump

Variable	Mean	Std	Q1	Q3	skewness
Clinton	-0.8508765	0.75633	-1.35681	-0.4070893	0.4685977
Trump	0.6101515	0.7304526	0.1985637	1.106471	-0.6183252

The histogram for Clinton and Trump is shown in Figure 4 and Figure 5

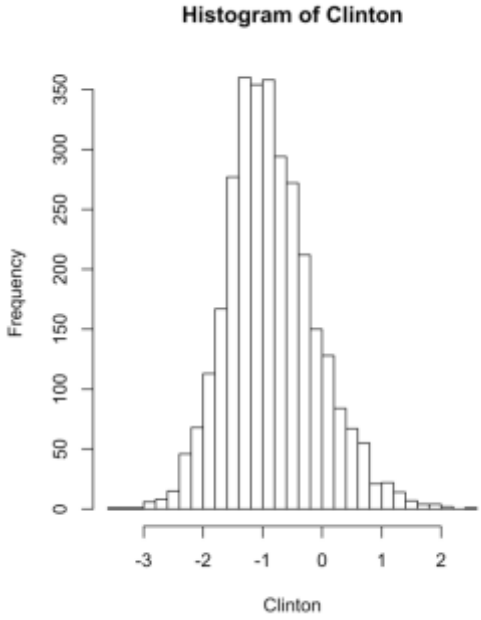

Figure 4 histogram for Clinton

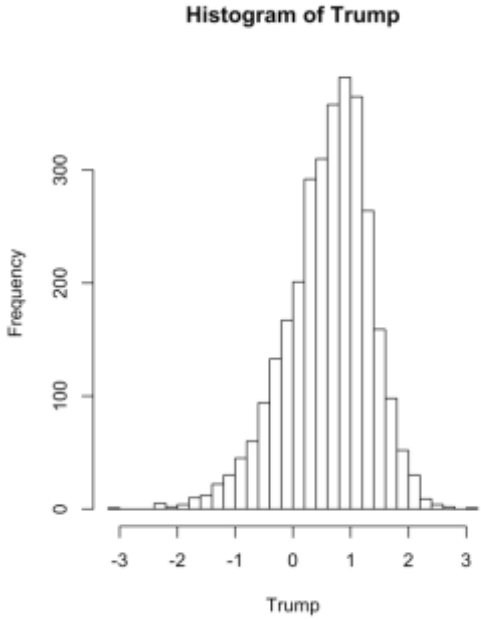

Figure 5 histogram for Trump

After the transformation, the skewness of Clinton is 0.4685977, and the skewness of Trump is -0.6183252. The skewness is significantly reduced by the transformation.

We draw the scatter plot between Clinton, Trump and the demographic and economic characteristics variables using R. See Figure 6 and Figure 7.

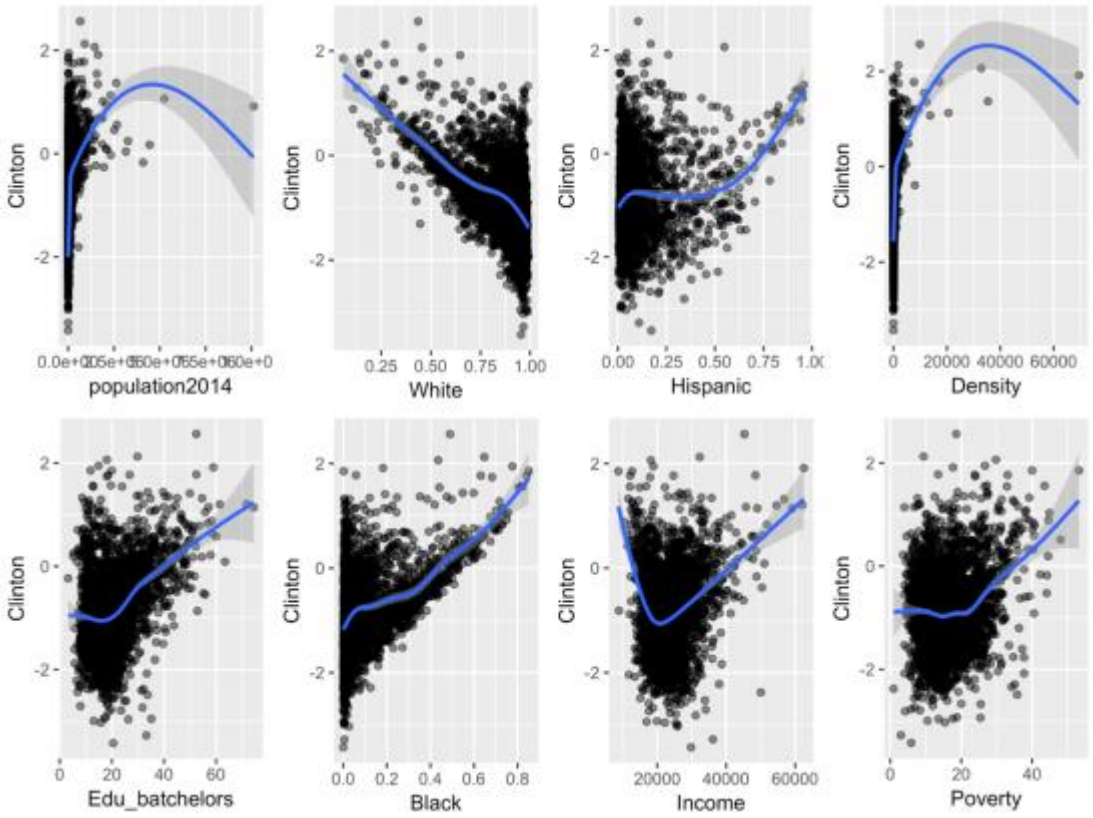

Figure 6 Scatter plot between Clinton and the demographic and economic characteristics variables

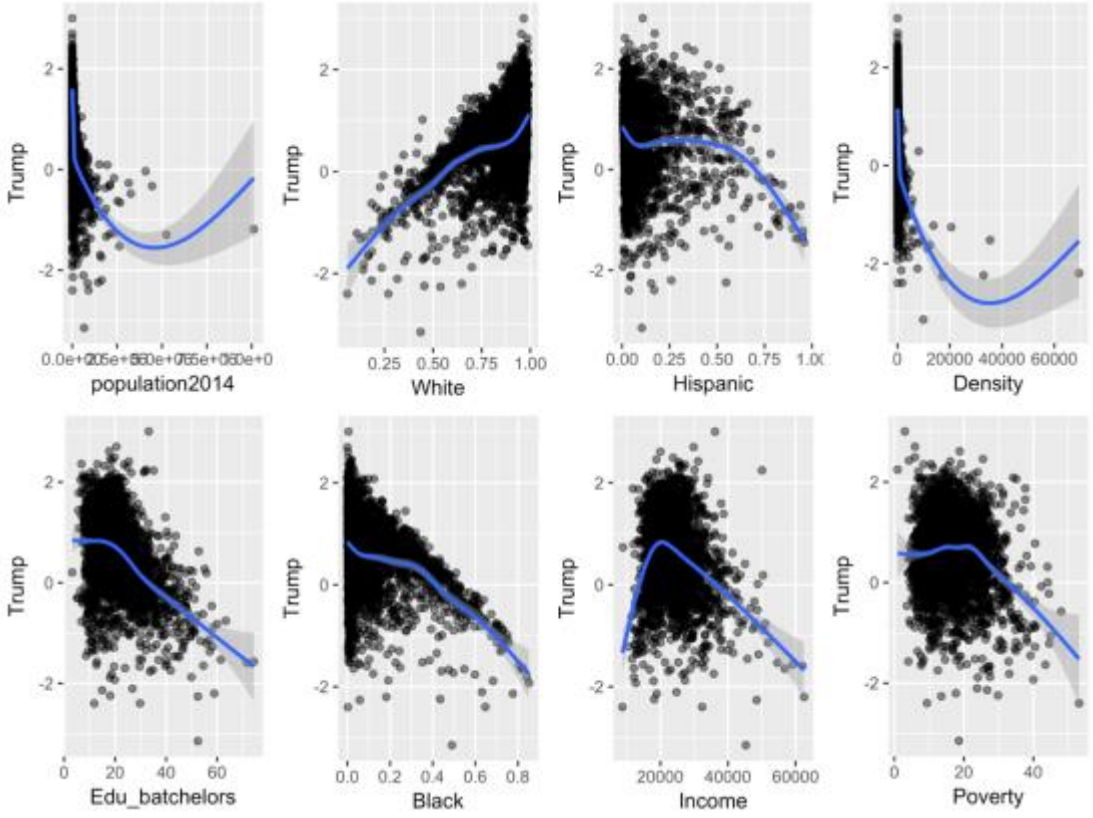

Figure 7 Scatter plot between Clinton and the demographic and economic characteristics variables

We can see from the scatter plot

1. The curve of Clinton is the opposite of the curve of Trump, which is in line with reality.
2. Clinton and Trump is greatly influenced by demographic characteristics and economic characteristics, we will follow it with quantitative correlation analysis.
3. Variables White, Edu_batchelors, Black is linearly correlated with variables Clinton and Trump.
4. Population, Density, Income, Poverty data points concentrated in an area and are unevenly distributed, and therefore conclusions cannot be drawn from the scatter plot.

We draw box plots to explore the relationship between Clinton, Trump and Region ,Sub.Region. See Figure 8 and Figure 9.

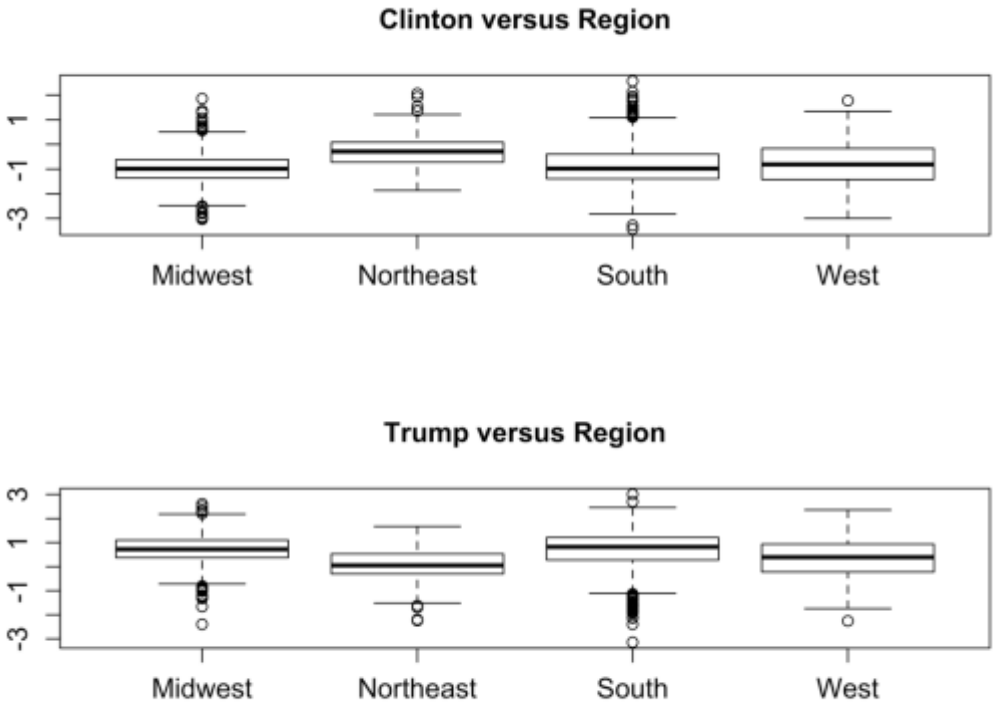

Figure 8 Box plots for Clinton, Trump versus Region

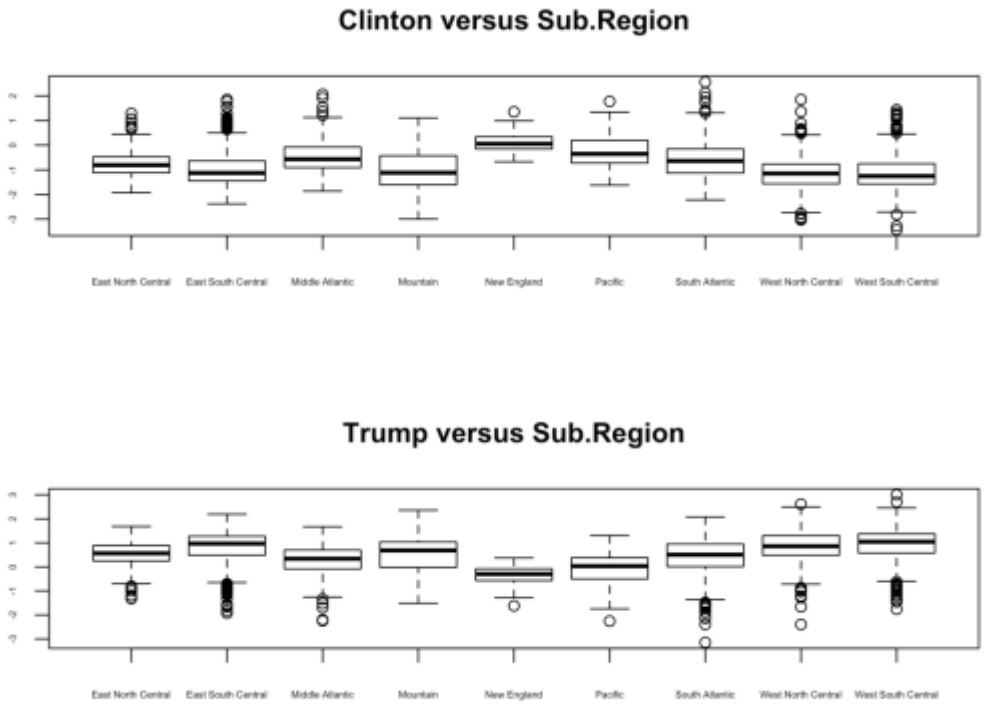

Figure 8 Box plots for Clinton, Trump versus Sub.Region

The above two figures tells that the influence of geographical areas to the voters' voting behavior cannot be ignored, therefore we consider adding Region and Sub.Region as random effects to the mixed effects model.

3. Correlation Analysis and Variable Selection

We conduct Pearson and Spearman correlation analysis to further explore the relationship between voters' behavior(poll results) and the demographic and economic characteristics of the county.
Poll results data includes Clinton, Trump, Romney, Obama (after transformation).
The demographic and economic characteristics includes population2014, Edu_bachelors, White, Black, Hispanic, Income, Density, Poverty.
The results are as follows,

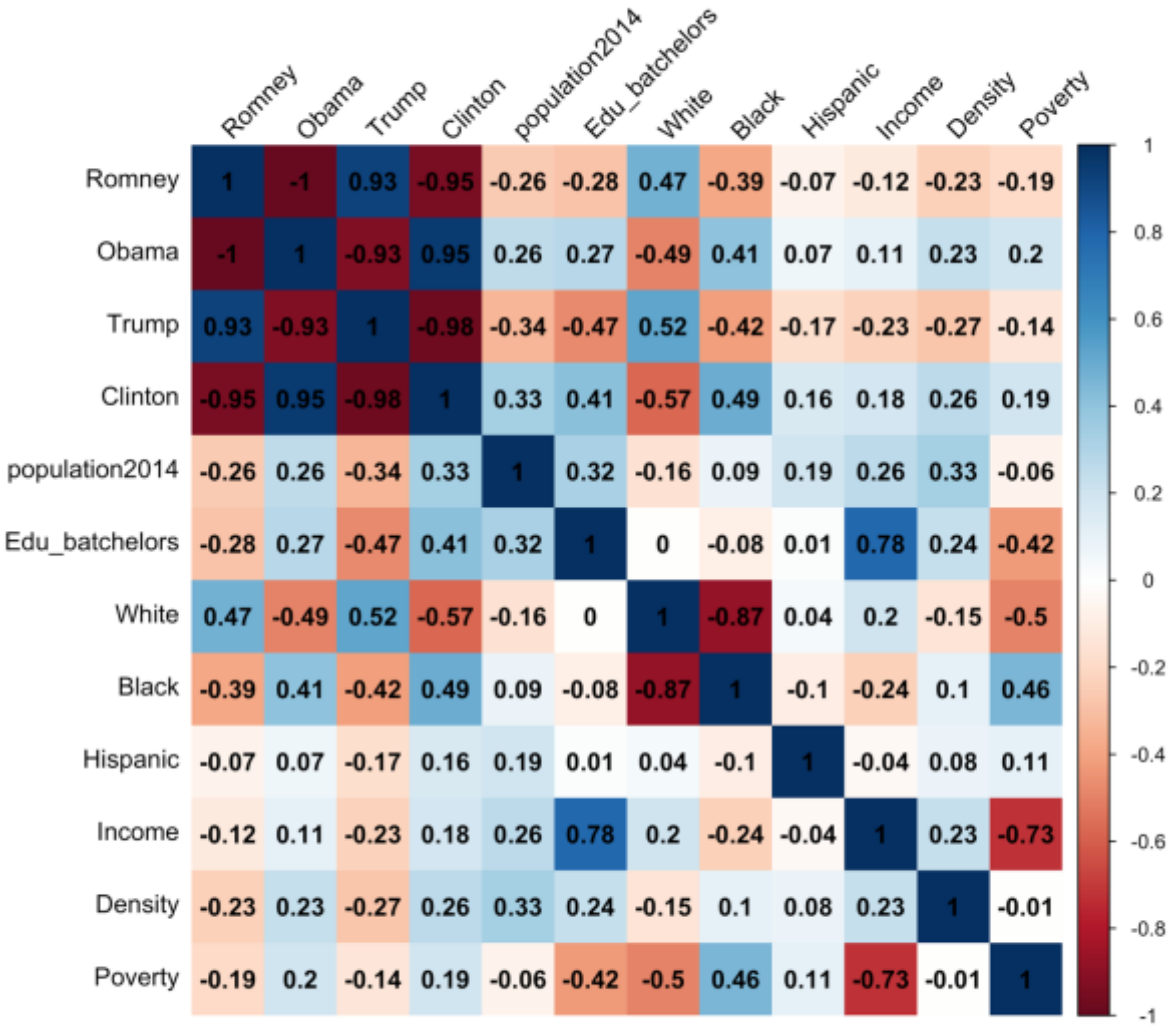

Figure 9 Pearson correlation coefficients between poll results and the demographic and economic characteristic

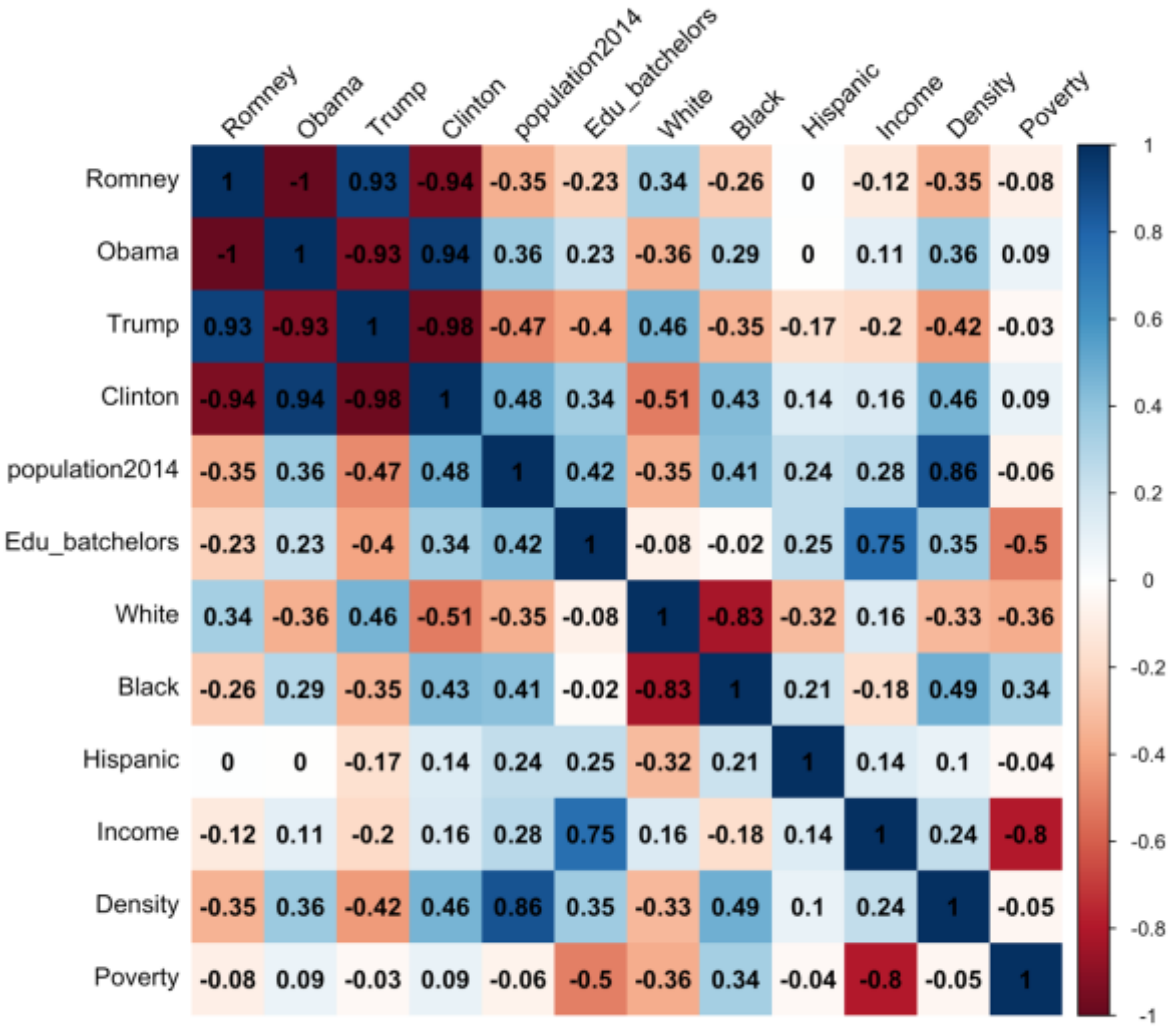

Figure 10 Spearman correlation coefficients between poll results and the demographic and economic characteristic

From the results of correlation analysis we can draw the following conclusions:

1. There is not much difference between the results of Spearman correlation analysis and Pearson correlation analysis. We can apply regression next to examine the quantitative relationship.

2. From the 4 top left grids of the correlation analysis results, We can see that the two candidates from Democratic or Republican are highly positively correlated. The positive correlation corresponds to the feature of the US election.

3. Regarding the strength of the correlation between Clinton, Trump and the demographic and economic characteristic, We can conclude that both Clinton and Trump are highly correlated with population, Edu_batchelors, White, Black , Density, and moderately correlated with Hispanic, Income, and weakly correlated with Poverty.

4. Regarding the direction of the linear relationship, Hillary Clinton get relatively more votes from more populous and more dense counties . Also, counties with relatively high proportion of blacks, Hispanics, Latinos and of high average income will vote for Clinton. Trump, on the other hand, was the exact opposite of Hillary Clinton, who won a relatively large number of votes in less populated and sparse counties and was welcomed by whites.

5. It is worth noting that there exist moderate or high correlation between (Edu_batchelors, Income, Poverty) and (White, Black).

6. Based on the above information, demographic characteristics and economic characteristics have a strong influence on the poll results of the 2016 US presidential election. In the regression analysis that follows, we consider remove variable income to obtain better model and reduce collinearity.

5. Multivariate Regression Model and Fixed Effects Model

The results of the four regression model are reported in Table 5.

Model 1 is a multivariate regression model that includes all the economic and demographic variables that are population2014,Edu_batchelor, White, Black, Hispanic, Income, Density, Poverty. The results of the model indicate that all the economic and demographic variables are statistically significant.

Model 2 is a multivariate regression model that includes all the economic and demographic variables except income. The results of the model indicate that all the economic and demographic variables included are statistically significant.

The model 3 and model 4 are mixed effects model which include Region and Sub.Region as random effects and economic and demographic variables as fixed effects. The mixed effects model indicate that while economic and demographic characteristics affect the voters' behavior, the geographic location also affect the behavior. The reason is that the observed economic and demographic characteristics cannot fully account for the poll results, and that the unobserved region-specific cultural characteristics play a part in determining poll results.

Model 3 is mixed effects model which include Region as random effects and economic and demographic variables as fixed effects. The results of the model indicate that population2014,Edu_batchelor, White, Black, Hispanic, Poverty are statistically significant. Specially, Density that are previously found to be significant, now is found to be insignificant. The result suggest that Region includes most of the information in Density, and Density becomes less important in explaining county-level pattern in 2016 US election.

Model 4 is mixed effects model which include Sub.Region as random effects and economic and demographic variables as fixed effects. The results of the model indicate that Edu_batchelor, White, Black, Hispanic, Poverty are statistically significant. Compared to model 3, population2014 that are previously found to be significant, now is found to be insignificant. The result suggest that Sub.Region includes most of the information in population2014, and population2014 becomes less important in explaining county-level pattern in 2016 US election.

It should be noted that geographic location factors has already included the information contained in population2014 and Density. If we would like to investigate the relationship between poll results and population, density, we should look into the results in multivariate models.

Table 5 Results of the four regression model

Regression Results				
Variables	Model 1	Model 2	Model 3	Model 4
population2014	0.58886a	0.059808a	0.042992a	0.013411
	(0.10277)	(0.010280)	(0.009358)	(0.008629)
Edu_batchelors	0.58886a	0.326768a	0.277318a	0.270765a
	(0.015731)	(0.011051)	(0.014454)	(0.013262)
White	-0.264481a	-0.266597a	-0.251909a	-0.292513a
	(0.020045)	(0.020047)	(0.018512)	(0.017190)
Black	0.133317a	0.133273a	0.233107a	0.18063a
	(0.018993)	(0.019010)	(0.018420)	(0.017137)
Hispanic	0.120022a	0.120001a	0.16864a	0.222836a
	(0.00948)	(0.009489)	(0.009079)	(0.009027)
Income	0.054343a		0.063268a	0.046034a
	(0.02094)		(0.019099)	(0.017439)
Density	0.034167a	0.03921a	0.009443	0.013505a
	(0.009965)	(0.009783)	(0.009143)	(0.008360)
Poverty	0.111117a	0.082919a	0.16299a	0.131398a
	(0.016277)	(0.012130)	(0.014988)	(0.013825)
Region				
			(0.07743)	
Sub.Region				
				(0.1560)
Adjusted R square	0.5542	0.5534		
F-statistic	484.5a	551.8a		
Residual Standard Error	0.505	0.5054		
Deviance			4001.4	3430.5
AIC	4589.806	4594.553	4023.412	3452.5
BIC	4650.236	4648.94	4089.885	3518.9

[a,b]denote significance at the .01 and .05 levels, respectively

The residuals of the four models are shown in Figure 11,

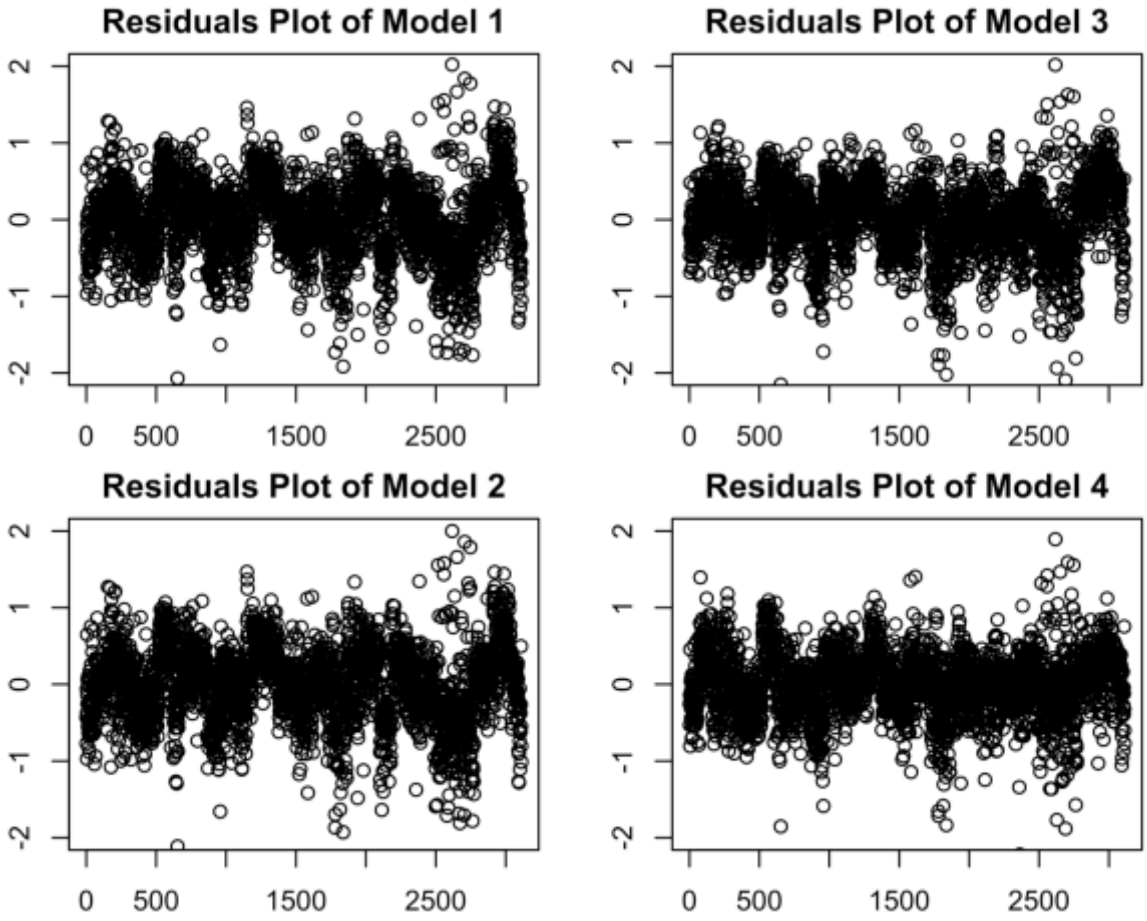

Figure 11 Residuals of the four models

The residuals plot tells us that the residuals of any of the four models follow the same distribution and is iid with mean equal to 0. The results is in accord with the assumption of the regression model.

We can draw the following conclusion from the multivariate regression model and mixed effects model,

1. The coefficients of all the economic and demographic variables except Density and population2014 is the same in all of the four models. We can conclude that percent of a county's vote that was cast for Clinton increased as the percent of the county's adult population that graduated from college increased, as the percent of the county's population that was black increased, as the percent of the county's population that was Hispanic increased, as the average income of the county's increased, as the percent of the county's population that was below poverty line increased, and as the percent of the county's population that was white decreased.

2. The coefficients of Density and population2014 is the same in the multivariate regression models. We can conclude that percent of a county's vote that was cast for Clinton increased as the population of the increased, and as the density of the county

increased.

3. Model 1 and model 2 are both multivariate regression models. Model 2 take out the variable income which is included in model 1. The coefficients of the independent variables in the results of model 2 is not much different from those of model 1. We can conclude that the collinearity between income and other economic and demographic variables don't have a large impact on the results of the regression. And also, adjusted R square of model 1 and model 2 is 0.5542 and 0.5534, respectively. AIC and BIC of model 1 and model 2 are 4589.806, 4594.553, and 4650.236, 4648.94, respectively. We can conclude that model 1 is better than model 2 and should be kept.

Table 6 Coefficient of economic and demographic variables and the impact of Region(intercept) in model 3

Coefficients of economic and demographic variables and the impact of Region(intercept) in model 3									
	(Intercept)	population2014	Edu_batchelors	White	Black	Hispanic	Income	Density	Poverty
Midwest	-0.64	0.043	0.277	-0.252	0.233	0.169	0.063	0.009	0.163
Northeast	-0.332	0.043	0.277	-0.252	0.233	0.169	0.063	0.009	0.163
South	-1.094	0.043	0.277	-0.252	0.233	0.169	0.063	0.009	0.163
West	-0.825	0.043	0.277	-0.252	0.233	0.169	0.063	0.009	0.163

Table 7 Coefficient of economic and demographic variables and the impact of Sub.Region(intercept) in model 4

Coefficients of economic and demographic variables and the impact of Sub.Region(intercept) in model 4									
	(Intercept)	population2014	Edu_batchelors	White	Black	Hispanic	Income	Density	Poverty
East North Central	-0.404	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
East South Central	-0.998	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
Middle Atlantic	-0.467	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
Mountain	-1.046	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
New England	0.033	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
Pacific	-0.511	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
South Atlantic	-0.93	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
West North Central	-0.807	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131
West South Central	-1.341	0.013	0.271	-0.293	0.181	0.223	0.046	0.014	0.131

4. The model 3 and model 4 are mixed effects model which include Region and Sub.Region as random effects and economic and demographic variables as fixed effects. Table 7 and Table 8 show the coefficients of economic and demographic variables and the impact of Region(intercept) in model 3 and model 4. It should be noted that the coefficients of the economic and demographic variables is same given different Region or Sub.Region in the same model. Because the coefficients of economic and demographic variables are fixed effects, they will not change even if the random effects change. From the intercept shown in the tables, we can conclude that Clinton is more popular in Northeast, and less in West and South. In term of sub-region, we can conclude that the percent of a county's vote that was cast for Clinton is distinctly larger in New England, Middle Atlantic, and distinctly smaller in West South Central, East South

Central, and Mountain. The deviance for model 3 and model 4 are 4001.4, 3430.5, respectively. AIC and BIC of model 1 and model 2 are 4023.412, 4089.885and 3452.5, 3518.9, respectively. The mixed effects model outperform multivariate model, and model 4 outperform model 3.

5. Due to the collinearity between Black and White, we tried to build a model without either of the two. However, it turn out the performance of the model is much worse than the existing one. Therefore, we decided to discard such models.

6. It should be noted that the results pertaining to the effect of region-specific cultural characteristics must be interpreted somewhat cautiously. If the county-level demographic and economic variables included in the regression models properly controlled for all the demographic and economic characteristics that influenced the county' s voting behavior, the fixed effects capture the variation in the percent of the vote cast for Gore that was caused by region-specific cultural characteristics. If, however, relevant demographic and economic characteristics that affect voting behavior are omitted from the models, the fixed effects capture variation that was attributable to not only regional-specific cultural characteristics but to the omitted factors as well.