Technical Report:
A Spatial Population Study of the Counties in the 2016 Presidential Elections

Dr. Raid Amin and Shawn Harrel

Department of Mathematics and Statistics,
University of West Florida
ramin@uwf.edu

April 2017

Abstract

Using a multitude of covariates, we us statistical procedures to select the most significant variables. Poisson regression was used for modeling county counts for Trump based on population. Furthermore, we use logistical regression as well as multiple linear regression to create adequate models for both predicting how many votes a county can expect to get for a political party and for profiling a county. Using SaTScan we developed clusters that show where certain covariates may be higher compared to the rest of the contiguous counties in the U.S. Multivariate analysis was done to identify grouping of covariates. Finally, we utilized factor analysis to identify three factors and how our covariates are correlated to them.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Problem Statement

The United States of American has been holding presidential elections since 1789. Since 1789, 56 elections have taken place, where 52 out of the 56 elections had a winner who won both the popular vote and the Electoral College vote. However, this election was not in the set of the 52 previous elections, it was instead, one of the 4 elections where the winner did not win the popular vote. But, instead won the Electoral College vote.

In the spring of 2015, two candidates, Hillary Clinton and Donald Trump announced their intentions to run for the President of the United States of America. During the run for Presidency, Hillary Clinton was projected to win both the Electoral college vote and the popular vote [12]. As we know, the outcome was not as predicted.

Using a slew of statistical methods, we wish to profile all the contiguous counties in the U.S. These methods include Poisson regression, multiple linear regression, logistic regression, multivariate analysis, and factor analysis. As well as, comparison of means for each covariate where we compare the means of the political parties, i.e. is the mean for obesity of democratic counties different from the mean of obesity of republican counties. Utilizing these statistical methods, we wish to establish an accurate model for profiling each county in the continental U.S.

1.2 Literature Review

According to [13], they state that both economic and non-economic factors can be used in a multiple linear regression model to predict the 2012 U.S. Presidential election. They establish a 95% confidence interval of (51.818, 54.239). Furthermore, their model can “comfortably” predict the democratic parties 2012 election, as well as the 2008 presidential election successfully.

Hsieh et al. show that they can adjust the required sample size for a multiple logistic regression model by a variance inflation factor [5]. Furthermore, their method requires no assumptions of low response probability in the logistic model [5]. They go on to show that they can derive the variance inflation factor for linear regression model, and through computer simulation, they are able to show that the same variance inflation factor applies to the logistic regression model with binary covariates [5].

Greenwald et al. used two implicit and two self-reporting measures of racial preference for European American relative to African Americans measure their symbolic racism and political conservatism [4] for predicting the 2008 U.S. Presidential election. They used logistic regression for measuring prediction of votes by race attitude measures. They used multiple regression for predicting the relation of race attitude and how it measures to conservatism. Furthermore, they used multivariate model for prediction of voting intention [4].

Chapter 2

Data and Software

2.1 Data

The data used in this research was collected from five different sources. These sources include County Health Rankings, the Environmental Protection Agency, the U.S. Census Bureau, the U.S. Religion Census, and CNN. Given these five sources, over 20 covariates were used for a variation analyses. All covariates that were transformed to normal quantiles were done using Blom's method.

2.1.1 Limitations and Data Cleaning

When data was obtained for voting counts based on counties, the information for Alaska was not present. That is, since information about the presidential election for Alaska is not present it is omitted from this research. Furthermore, Hawaii was also omitted from this study, thus we will only focus on the contiguous counties in the U.S. Also, if a data source included protectorate states, these data points were also removed. The population density is based off the 2014 population estimates; thus, the population density may have changed slightly over the past two years for all counties. Furthermore, the data used for particulate matter 2.5 ($PM_{2.5}$) and National Air Toxics Assessment (NATA) is from 2011. Therefore, the values used for $PM_{2.5}$ and NATA will not be current.

It was not until all analyses had been completed and data put into SaTScan as well as ArcGIS that a single counties data was missing. This may be due to a conflict of FIPs between my data set and the data for FIPs ArcGIS uses. Therefore, all maps will not include this single county. However, all analyses ran had this missing county included.

Not every county will provide their data with respect to some variable. When this is the case, data imputations need to be done. The two data sources required data to be imputed upon U.S. religion Census and the County Health Rankings. When imputing for any missing county value, the mean of the respective state was used. There were two variables that required this more than the others, they are the rate for high school graduation and the rate for violent crimes per county.

2.1.2 Variable Definition

The following variables used in this research came from county health rankings website, however these variables are not well defined. Adult smokers are classified as someone who smokes every day or most days of the week and has smoked at least one hundred cigarettes in their lifetime according to county health rankings [1]. Adult obesity as defined by the county health ranking as a person over the age of twenty whom has a body mass index greater than or equal to 30 kg/m^2 [1]. Someone who is classified as an excessive drinker consumes more than 4 (women) or 5 (men) drinks on a single occasion in the past 30 days; or, more than 1 (women) or 2 (men) drinks per day on average [1]. For high school graduation, this variable is based off the ninth-grade cohorts that graduate high school in four years [1]. The variable “some college" is based off the population from the ages 25–44 with some post–secondary education [1]. As far as unemployment, this variable is the total unemployed persons, as a percentage of the civilian labor force ages greater than or equal to 16 [1]. Violent crime is composed of four offenses: murder and non-negligent manslaughter, rape, robbery, and aggravated assault [1]. Mental distress is defined by county health rankings as an adult who suffers from mental health, which includes stress, depression, and problems with emotions at least 14 days out of a 30-day period [1]. Insufficient sleep is defined as an adult who sleeps less than 7 hours per night [1]. The median household income is defined as the income where half of households in a county earn more than the other half of household [1].

The variables used from the Environmental Protection Agency are $PM_{2.5}$ and NATA. NATA is a national scale screening analysis of air toxic emissions [3]. Particulate matter is a complex mixture of extremely small particles and liquid droplets that are in the air, where the size of the particles are approximately 2.5 microns or less across [3].

2.2 Software

2.2.1 SAS

SAS allows any users to improve data delivery, analysis, reporting, data movement across a company, data mining, forecasting, statistical analysis, and more [11]. Originally, SAS was used for statistical analysis, however the software has become robust and handles a multitude of tasks. Throughout the study, SAS was used for normalizing data using Blom's method. During this study SAS was also used for combining multiple data sets, data cleaning, data imputations, running regression analysis, factor analysis, and correlation matrices.

2.2.2 SaTScan

SaTScan is a software that was developed by Dr. Martin Kuldorff. SaTScan is used widely in public health and other fields to identify high or low clusters of illness or other events across space and time. SaTScan allows for a multitude of analyses, within each analysis multiple ways to run the analysis are available. SaTScan provides output that can be used in Excel, Google Earth, as well as ArcGIS.

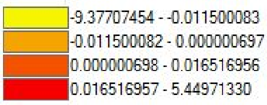
All spatial analyses used in this study had a maximum spatial cluster size for 10% of the population at risk. Furthermore, only we only scanned for areas with high rates, that is SaTScan searched for values larger than the mean of the continuous U.S. for said analysis. SaTScan outputs cluster in a hierarchical manner, that is suppose there are clusters from $1,2,3,\ldots,n$, then cluster 1 is considered the most likely cluster, cluster 2 is considered the second most likely cluster, and so on. During the duration of this study all covariates used in SaTScan were first normalized before any clusters were identified. Also, all analyses done used 999 Monte Carlo replications.

2.2.3 ArcGIS

There are several ways to access ArcGIS, one of which is ArcGIS Online, the other is a desktop program. Using ArcGIS as a desktop program, you can use and create maps and scenes, access ready-to-use maps, layers and analytics, publish data as web layers, as well as collaborate and share [2].

ArcGIS was used in collaboration with SaTScan, as well. All clusters that were established in SaTScan were then given a geographical representation in ArcGIS. From there the maps were layered and a graduated color was assigned based on assigned bounds. All covariates used in ArcGIS were transformed to standard normal. They were then mapped as quantiles with three levels for graduated color. Apart from of a select few maps, all maps were adjusted for population density. Furthermore, the legends use a graduated color system from yellow to red. Yellow represents low values, or the bottom half of a normal curve; whereas red represents high values, or the top half of a normal curve. The color ramp is broken up into classes, where each class contains an even amount of observations. That is, suppose we have the following color ramp

Fig. 2.1 Color Ramp



🟨 -9.37707454 - -0.011500083

🟧 -0.011500082 - 0.000000697

🟧 0.000000698 - 0.016516956

🟥 0.016516957 - 5.44971330

Then each of these four class represent 25% of the data.

Chapter 3

Analyses and Results

3.1 Regression

Regression is among one of the most widely used techniques for analyzing multiple variables, as well as factors. Regression analysis is used for modeling a relationship between a response variable, often thought of as y_i, and regressors, often thought of as some x_i, where $i = 1, 2, 3, \cdots, n$. The power of regression comes from its ability to explain $\hat{y}$ from x, given the model is built appropriately and x is in the range used in the data. There are a multitude of regression models that can be utilized such as polynomial regression, nonlinear regression, nonparametric regression, ridge regression, etc. Though there are many different types of regression we will be using Poisson regression, multiple linear regression, and logistic regression throughout this study.

3.1.1 Poisson Regression

Poisson regression is used to for modeling count data, that is the response variable y_i consists of count data and is considered a rare event. Furthermore, the independent variables are continuous. The Poisson model given as a general linear model for count data is illustrated as the following

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

with the following assumptions: the response variable has a Poisson distribution, where the expected count of y_i is $\mathbf{E}(\mathbf{Y}) = \mu$, and any set of $X = (X_1, X_2, \cdots, X_k)$ are explanatory variables.

The values obtained for Wald Chi-Square are significant for identifying which covariate plays the largest role in the model. The larger the value is for Wald Chi-Square the more that variable contributes to the model. Conversely, the smaller the Wald Chi-Square value is the less the variable adds to the model.

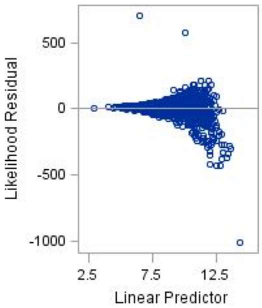
Using the voting counts for Trump for each county as the response variable and the population density for 2014 as the independent variable, we get the following

Table 3.1 Poisson Parameter Estimates

Parameter	Estimate	Wald Chi-Square
Intercept	-1.5297	1.086E8
Population Density	-0.0001	1597063

Table 3.1 illustrates that the model for predicting Trump votes per county is $g(\mu) = -1.5297 + -0.0001 x_{PopulationDensity}$. The model indicates that the lower the population density the less votes Trump gets. Furthermore, we get a model where the residuals are heteroscedastic. That is, the variability of the variable, voting counts for Trump, is unequal across the range of values for variable population that predicts it. The comparison of the likelihood residuals versus the predictor value, which illustrates heteroscedasticity can be observed in figure 3.1.

Fig. 3.1 Likelihood Residuals versus Predictor



The log likelihood ratio is used to find a p-value for clusters pertaining to a spatial scan in SaTScan. According to [6], the Log Likelihood Ratio is given as:

$$LLR(z) = \left(\frac{c}{E[c]}\right)^{c} \left(\frac{C-c}{C-E[c]}\right)^{C-c} I()$$

where C is the total number of cases, c is the observed number of cases, $E[c]$ is the covariate adjusted expected number of cases, and I() is the indicator function.

$$I() = \begin{cases} 1 & \text{When the window has more cases than expected} \\ 0 & \text{Otherwise} \end{cases}$$

Fig. 3.2 Trump Counts Per County Based off 2014 Population

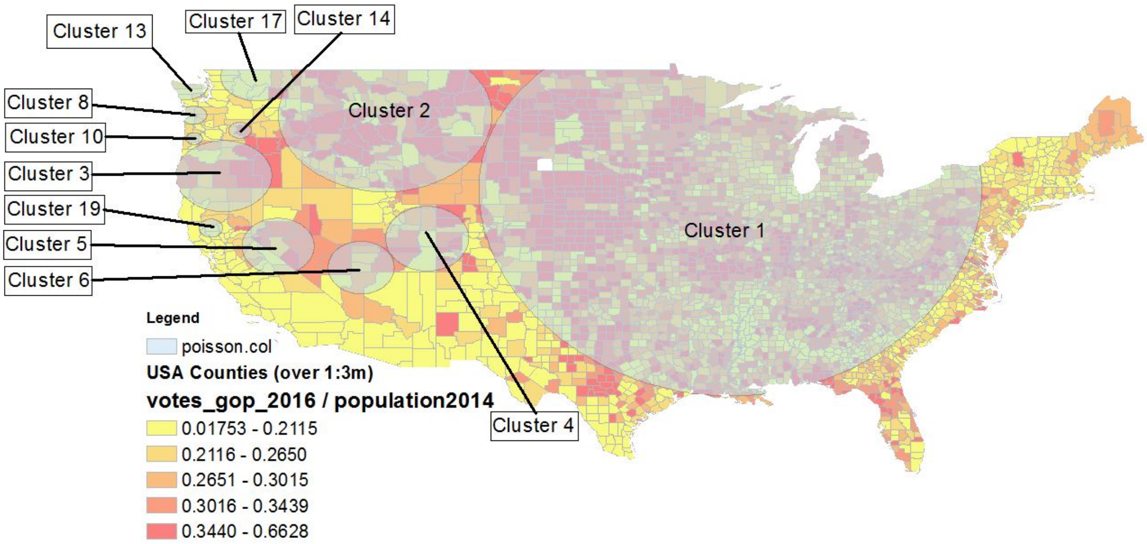


Figure 3.2 is a map of the votes Trump received normalized with respect to the 2014 population. Furthermore, we can detect that Trump received more votes in counties with lower population. We can also observe that there are 12 clusters that are all perceived as significant. The relative risk for each cluster is given in table 3.2.

Table 3.2 Relative Risk Per Cluster

Cluster	Relative Risk	P-value
1	1.48	<0.001
2	1.32	<0.001
3	1.33	<0.001
4	1.52	<0.001
5	1.45	<0.001
6	1.36	<0.001
8	1.25	<0.001
10	1.19	<0.001
13	1.13	<0.001
14	1.30	<0.001
17	1.07	<0.001
19	1.12	<0.001

The relative risk is used to identify the probability of an event occurring. Furthermore, table 3.2 illustrates the cluster outputted in a descending order of relative risk. Hence, relative risk indicates that the larger the value, the more likely that cluster is.

3.1.2 Logistic Regression

Since we are testing the outcome of the Presidential Election for 2016 with only two candidates Hillary Clinton and President Donald Trump implies that our response variable is binary. This leads into logistic regression models, that is, models with binary response variables. The goal of logistic regression is to develop a superlative model so that we can conjecture a relationship between the response variable and the independent variables. The following model is given by:

$$y_i = x_i'\beta + \varepsilon_i$$

where $x_i' = [1, x_{i1}, x_{i2}, \cdots, x_{ik}]$, $\beta' = [\beta_0, \beta_1, \beta_2, \cdots, \beta_k]$ and the response variable y_i is either 0 or 1 [7]. Furthermore, the error term can be defined as

$$\varepsilon_i = \begin{cases} 1 - x_i'\beta & \text{if } y_i = 1 \\ -x_i'\beta & \text{if } y_i = 0 \end{cases}$$

Since the response variable is binary, we can further assume that it is a Bernoulli random variable. The Bernoulli random variable can be defined by the following probability density function:

$$y_i = \begin{cases} 1 & \text{if } P(y_i = 1) = \pi_i \\ 0 & \text{if } P(y_i = 0) = 1 - \pi_i \end{cases}$$

where $\pi_i = E(y_i) = x_i'\beta$.

For the remainder of the study $y_i = 1$ or event = 1 represents President Trump as the winner of a county and $y_i = 0$ or event = 0 represents Hillary Clinton as the winner of a county. That is,

$$y_i = \begin{cases} 1 & \text{Trump wins a county} \\ 0 & \text{Clinton wins a county} \end{cases}$$

A stepwise procedure was used to develop a logistic regression model of best fit. The purpose of using a stepwise procedure is to iteratively search for the most significant and least significant variables where the least significant variables are removed from the model.

Table 3.3 was ran using event = 1, that is, the following model is with respect to the counties President Trump has won.

The odds ratio, OR, indicates whether an association exists between the response variable and the independent variable. It is given as the following

$$Odds\,Ratio = \frac{odds(a)}{odds(b)}$$

If the OR is greater than one, then the variable is associated with higher odds of outcome. Whereas, if the OR is less than one, then the variable is associated with lower odds of outcome. If the value is one, then the variable does not affect the outcome.

Table 3.3 Stepwise Logistic Regression, Trump

Variable	Estimate	Odds Ratio	Wald Chi-Square
Intercept	7.7668	—	10.7677
Rate Obesity	0.2550	1.216	123.7755
Rate Excessive Drinking	-0.3741	0.634	87.9994
Rate Some College	-0.0744	0.928	51.3583
Rate Mentally Distressed	-0.5512	0.558	67.6335
Rate insufficient Sleep	0.0432	1.045	1.3887
Rate African America	-0.1155	0.891	148.9672
Rate Non-Hispanic White	0.0785	1.082	153.8193
Rate Female	-0.0901	0.918	5.3914
Rate Catholic	-0.0020	0.998	10.9405
Particulate Matter 2.5	0.1089	1.106	2.6371
NATA	0.0544	1.058	23.7159
Population Density 2014	-0.0010	0.999	23.6418

Based on the odds ratios we can identify Obesity, insufficient sleep, non-Hispanic White, particulate matter 2.5, and NATA as being associated with higher odds of outcome. Whereas, excessive drinking, some college, mentally distressed, African American, and Female are associated with lower odds of outcome. Finally, population density and catholic are extremely close to 1, which means these variables may not affect the outcome.

The variables that contribute the most to the model in order based off their Wald Chi-Square values are non-Hispanic Whites, African Americans, Obesity, excessive drinking, mentally distressed, and some college. Though these variables contribute greatly to the model, one must be observant of the estimate. That is, the rate of African Americans contributes greatly to the model base off the Wald Chi-Square value, but the estimate is negative. This implies that counties with a low African American population will likely vote for Trump. Conversely, counties with a high African American population will not likely vote for Trump.

Table 3.3 yields the following regression model given y is representative of Trump winning a county

$$\begin{aligned} y = 7.7668 &+ 0.2550x_{obesity} - 0.3741x_{drinking} - 0.0744x_{college} \\ &- 0.5512x_{distressed} + 0.0432x_{sleep} - 0.1155x_{AA} + 0.0785x_{white} \\ &- 0.0901x_{female} - 0.0020x_{catholic} + 0.1089x_{PM_{25}} + 0.0544x_{NATA} \\ &- 0.0010x_{density} \end{aligned}$$

The following variables where not considered significant enough to be put into the model. These variables include adult smoking, high school graduates, unemployment, median household income, population over 65, poverty, and violent crimes.

Using table 3.3, particularly the Wald Chi–Square values, we can determine which variables contribute the most to the model. That is, the larger the Chi–Square the more that variable contributes to the model. Looking strictly at our top three contributing variables, we have non-Hispanic White counties contribute the most to the model, where these counties are most likely to vote for Trump. The second most likely indicator on how counties voted is based on the rate of African Americans. Counties with large populations of African American were less likely to vote for trump. Obesity is the third largest contributor to how counties votes. That is, counties that are in general obese were more likely to vote Trump.

If we were to rerun our stepwise logistic regression model given y is representative of Clinton we would get the same variables, along with the same P-value for each variable, the same Wald Chi–Square and standard error for each variable. However, our slopes (estimate) for each variable will have a negative coefficient for the model that represents the Trump. That is, we have the following model given y is representative of Clinton winning a county

$$\begin{aligned} y = &-7.7668 - 0.2550x_{obesity} + 0.3741x_{drinking} + 0.0744x_{college} + 0.5512x_{distressed} \\ &- 0.0432x_{sleep} + 0.1155x_{AA} - 0.0785x_{white} + 0.0901x_{female} + 0.0020x_{catholic} \\ &- 0.1089x_{PM_{25}} - 0.0544x_{NATA} + 0.0010x_{density}. \end{aligned}$$

A Receiver Operating Characteristic Curve (ROC) is used to reflect the accuracy of the diagnostic test. That is, the ROC curve summarizes the performance between true positive and false positive [8]. The area under this curve is referred to as concordance index, where the concordance index is the traditional performance metric for a ROC curve [8]. Thus, the closer the concordance index is to one the better the prediction power is for the model.

Fig. 3.3 ROC Curve for All Model Building Steps

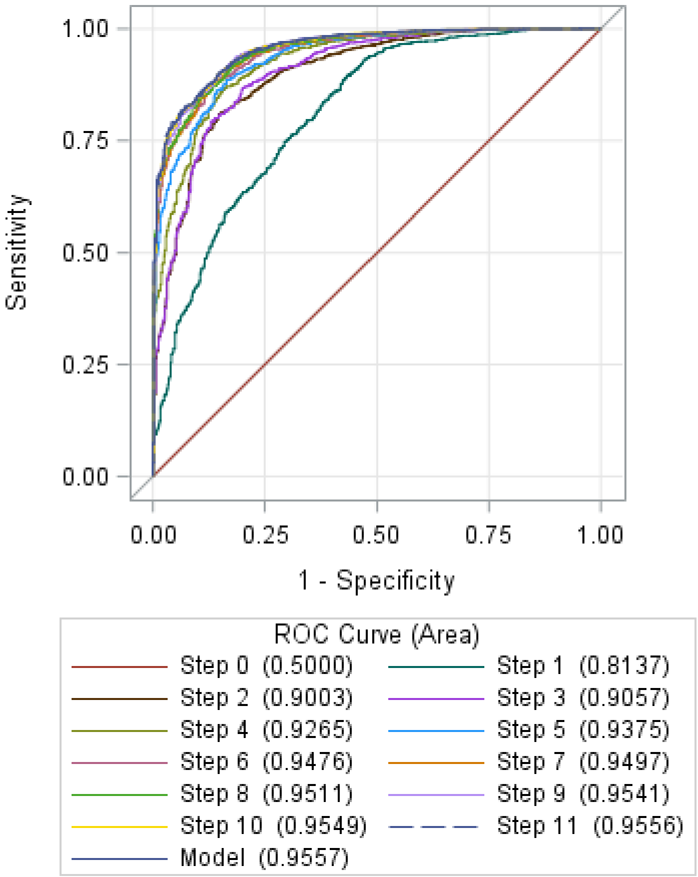


Figure 3.3 is an illustration of the ROC curve for our model using stepwise logistic regression. Step 0 indicates our first variable, which is non-Hispanic Whites, where the area or concordance index is 0.5000. The stepwise logistic regression goes through twelve iterations, where we arrive at our model. Furthermore, we can observe that as we step through the model the area under the curve strictly increases, which leads to a better model with each iteration. It can be identified that the final model has a concordance index of 0.9557, which indicates an excellent predictive power of the model. Therefore, from here on out our model will only include the following variables adult obesity, excessive drinking, some college, mentally distressed, insufficient sleep, African American, non-Hispanic White, female, Catholics, PM_{25}, NATA, and populations density for 2014.

The criterion for running the Bernoulli model in SaTScan is subject to a maximum spatial cluster size with 10 percent of the population at risk, the type of analysis is purely spatial, the model is Bernoulli, and the model scans for both low and high rates. Furthermore, the case file is the counts for Trump for each county and the control file contains the values 0 or 1, i.e. the response variable. When the model scans for low and high rates, this is indicative of the response variable, that is, the low rates indicate a 1 and high rates indicate a 0.

Fig. 3.4 DEM versus GOP

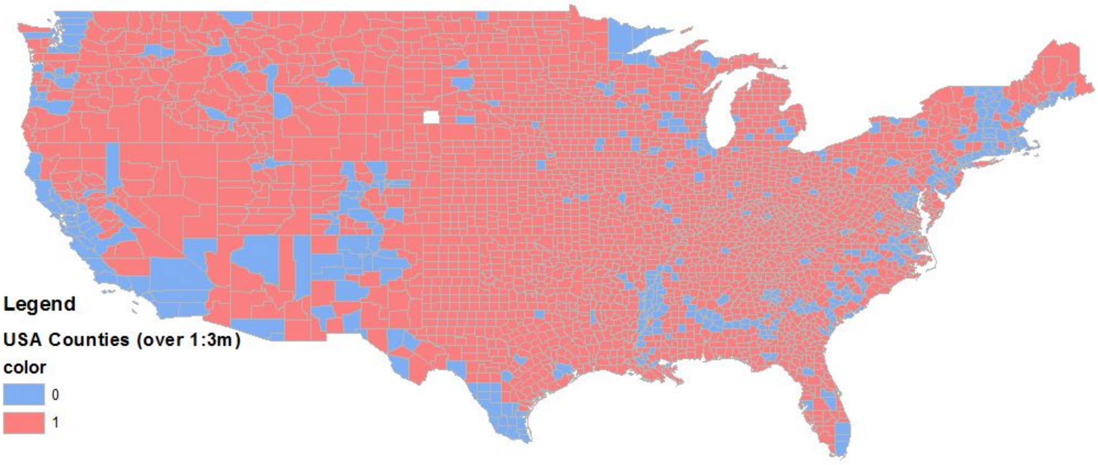


Figure 3.14 is an illustration of how the contiguous US voted by county. A blue county represents a win for the democratic party and a red represents a win for the republican part. Even though there are significantly more red counties than there are blue counties, counties that voted blue typically have a larger population. Which was illustrated using Poisson regression and observed in figure 3.2.

3.1.3 Multiple Linear Regression

Producing a multiple linear regression (MLR) model requires that our response variable is continuous. Thus, when we create a MLR model we are creating a model that with predict a continuous value. If we are to compare logistic regression to MLR, we can identify that logistic regression only predicts if a county will vote republican or democrat. Whereas, MLR will predict the percentage of votes a candidate will receive in said county. The multiple linear regression model is defined as

$$y = X\beta + \varepsilon$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

While working with multiple linear regression, our response variable is no longer binary, we instead make our response variable the ratio of $\frac{TrumpVotes}{TotalVotes}$ for each county. We then run our response variable against all our independent variables, which includes twenty covariates.

Table 3.4 shows a select number of models. Each model was selected based off the adjusted R^2 value, the Cp value, as well as its MSE. When selecting a model, we want to maximize adjusted R^2 value, minimize the Cp value, and minimize the MSE.

Even though the full model in table 3.4 has the best adjusted R^2 value, the lowest Cp, and the lowest MSE it is not necessarily good, there are a multitude of reasons for this. However, selecting a model with the least number of variables, while still having above reasonably good values for MSE, Cp, and adjusted R^2, may in fact be more appealing. Therefore, we will be selecting the model with twelve covariates for our multiple linear regression.

Table 3.4 Possible Multiple Linear Regression Models

Number in Model	Adj R-Square	Cp	MSE	Variables in Model
9	0.6904	147.6805	0.00754	Rate_Aobes Rate_excesDrinking Rate_someCollg Rate_menDistrss rate_Unins rate_AA rate_nonHisWhite rate_female NATA_RAW
10	0.6927	124.7937	0.00749	Rate_Aobes Rate_excesDrinking Rate_someCollg Rate_menDistrss rate_Unins rate_AA rate_nonHisWhite rate_female PovertyPercent NATA_RAW
11	0.7003	46.0913	0.00730	Rate_Aobes Rate_excesDrinking Rate_someCollg Rate_unemp Rate_menDistrss rate_Unins rate_AA rate_nonHisWhite rate_female NATA_RAW populationDensity14
12	0.7016	34.0396	0.00727	Rate_Aobes Rate_excesDrinking Rate_someCollg Rate_unemp Rate_menDistrss rate_Unins rate_AA rate_nonHisWhite rate_female PM_25 NATA_RAW populationDensity14
20	0.7036	21.0000	0.00722	Rate_Asmoking Rate_Aobes Rate_excesDrinking Rate_Hsgrad Rate_someCollg Rate_unemp Rate_VioCrime Rate_menDistrss rate_insuffSleep rate_Unins Median_HouseInc rate_ovr65 rate_AA rate_nonHisWhite rate_female PovertyPercent CATHRATE PM_25 NATA_RAW populationDensity14

When using logistic regression, it is difficult to determine if collinearity exists between the covariates. To determine that the covariates have a low variance inflation factor (VIF), multiple linear regression is used. According to [5], the VIF for the linear regression model shows to be the same VIF applied to the logistic regression model. Therefore, we will be using the VIF from our multiple linear regression model to indicate if any collinearity between our variables exists with respect to our logistic regression model. Furthermore, all analysis done on our logistic model in this section simply means that we are using the covariates identified in the logistic regression model, but a multiple linear regression analysis is used.

Multicollinearity exists when two variables that are in the model are correlated and can be explained by one another. Any model with a VIF above 20 is mildly multicollinear and any VIF above 30 is highly multicollinear. Our model yields the following VIF values

Table 3.5 Variance Inflation Factor for Both Models

Variable	VIF(Binary Model)	VIF (MLR Model)
Rate Obesity	1.6661	1.68401
Rate Excessive Drinking	2.2533	2.31676
Rate Some College	2.2389	2.60159
Rate Unemployed	—	2.04498
Rate Mentally Distressed	3.3678	3.61445
Rate Uninsured	—	2.45420
Rate insufficient Sleep	3.7397	—
Rate African America	2.9414	2.84736
Rate Non-Hispanic White	2.7539	2.84312
Rate Female	1.3591	1.33144
Rate Catholic	1.5703	—
Particulate Matter 2.5	1.5052	1.44054
NATA	1.8222	1.69028
Population Density 2014	1.5286	1.14562

Table 3.5 illustrates the values for the VIF for each variable in our model. The largest VIF identified for the logistic regression model is for the variable Insufficient Sleep, with a variance inflation of 3.7397. Whereas, the largest VIF identified for the MLR model is for the variable mentally distressed, with a VIF of 3.61445. We can conclude that multicollinearity does not exist in either of our models.

Table 3.6 Parameter Estimates for both Models

Variable	Estimate (MLR)	Estimate (LR)
Intercept	0.61477	1.36587
Rate_Aobes	0.00952	0.00867
Rate_excesDrinking	-0.00946	-0.01472
Rate_someCollg	-0.00245	-0.00434
Rate_unemp	-0.00827	—
Rate_menDistrss	-0.02540	-0.02815
Rate_uninsured	0.01044	—
Rate_insuffSleep	—	-0.00279
rate_AA	-0.00196	-0.00372
rate_nonHisWhite	0.00524	0.00310
rate_female	-0.00720	-0.00692
CathRate	—	-0.00012362
PM_25	0.00449	-0.00150
NATA_RAW	0.00185	0.00223
populationDensity14	-0.00000534	-0.00000749

Table 3.6 shows the parameter estimates for both models, where $\hat{y}$ is the of percent trump votes per county. We can further identify that the estimate for the MLR model excludes insufficient sleep and catholic rate and includes unemployment rate as well as uninsured rate. The converse can be said for the covariates of the logistic regression models'.

Fig. 3.5 Residual Histogram for the Logistic Model

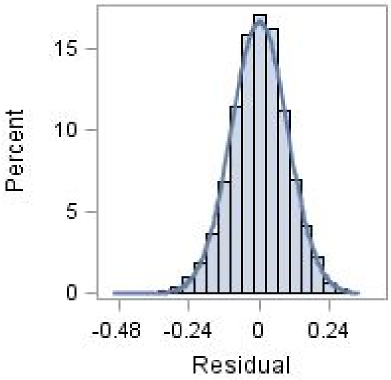
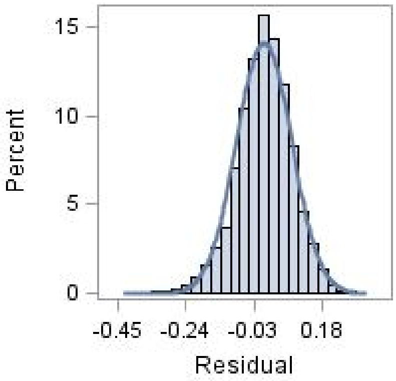


Fig. 3.6 Residual Histogram for the MLR model



Comparing figure 3.5 against figure 3.6, we can see the residuals for both histograms are approximately normal, however the residuals for the histogram using the logistic regressions covariates, fits the normal curve better than its counterpart.

Fig. 3.7 The Residual QQ-Plot for the Logistic Model

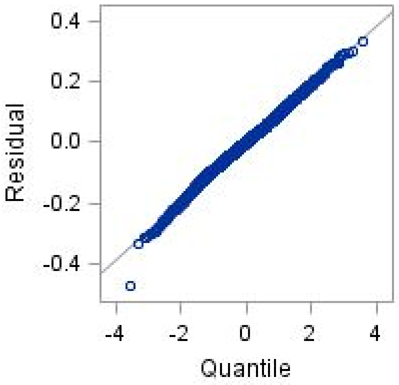
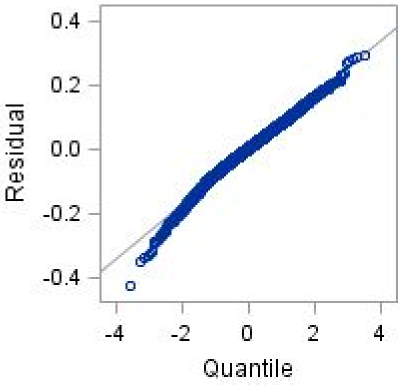


Fig. 3.8 The Residual QQ-Plot for the MLR model



Comparing figure 3.7 against figure 3.8, we can observe that the residuals for figure 3.7 have a better fit along the diagonal line compared to figure 3.8.

Based off [7], figures 3.5, 3.6, 3.7, and 3.8, indicate that the following assumptions are not violated

- The relationship between the response variable and the regressors is linear
- The error term ε has zero mean
- The error term ε has constant variance σ^2
- The errors are uncorrelated
- The errors are normally distributed

Table 3.7 Comparison of Models

Variable	MLR model	LR model
Adjusted R-Square	0.7016	0.6204
MSE	0.0073	0.0092
PRESS	23.0842	29.4359
CP	13	13

Even though the model we built using multiple linear regression show better results than the covariates established in the logistic regression model, the multiple linear regression model is of no consequence to us. This is because the MLR model cannot predict or profile how a county will vote, i.e. which candidate will win a county. The MLR model instead only predicts the percent votes Trump will get per county.

3.2 Full Model

Using single variable to establish a map for a model, figure 3.9 is an illustration of the contiguous counties in the US, where each county is a representation of normal quantiles pertaining to non-Hispanic Whites. That is, the data for non-Hispanic Whites has been converted to standard normal. Red represents a high concentration versus yellow represents a low concentration of non-Hispanic Whites. However, this map does account for the population with respect to each county. This leads way to isolating population density from the variable non-Hispanic Whites.

Fig. 3.9 Non-Hispanic White

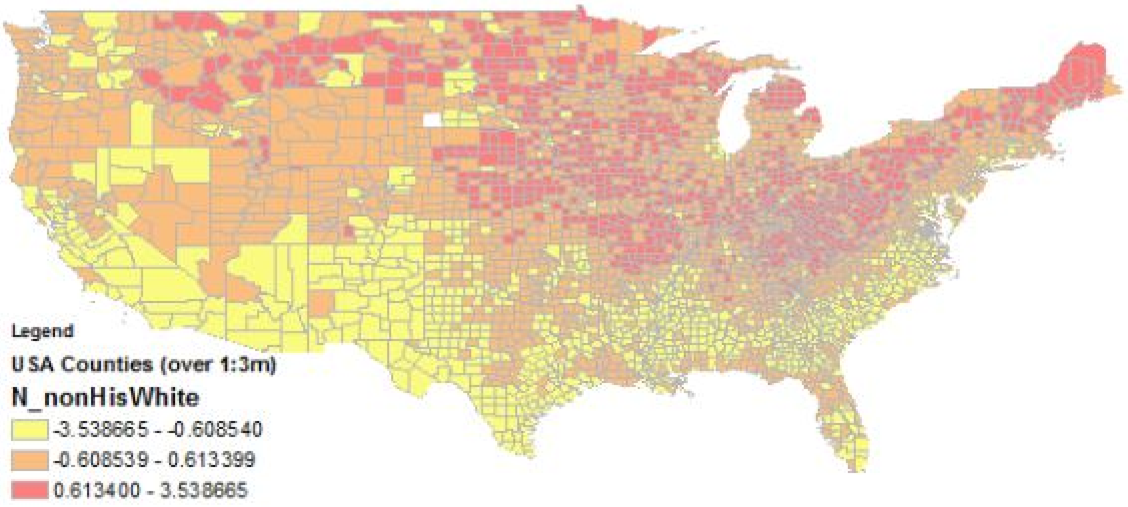
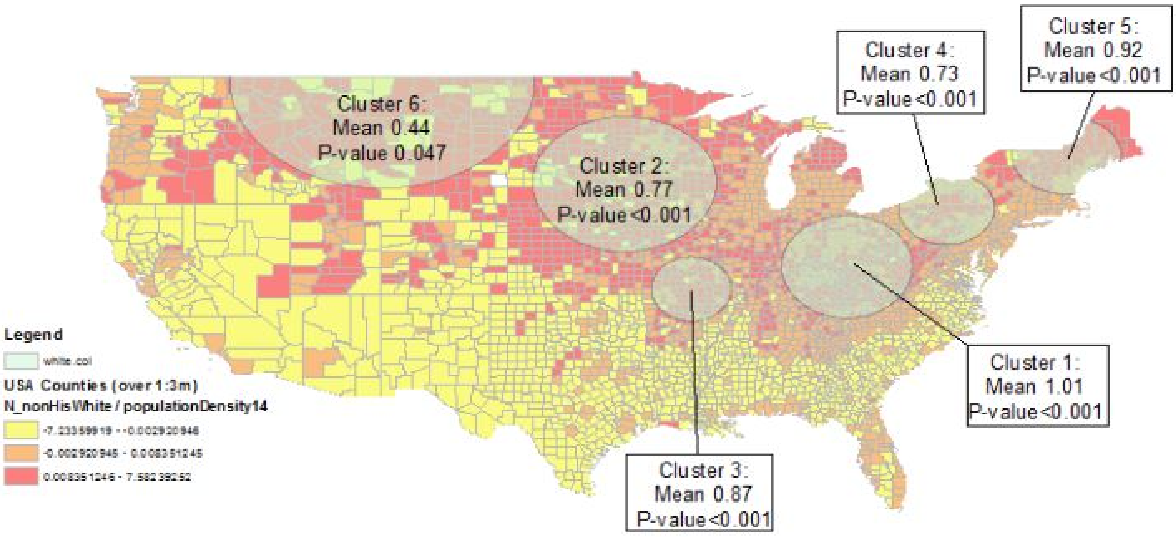


Fig. 3.10 Non-Hispanic White with Population Density Isolated



For figure 3.10 illustrates a map of non-Hispanic Whites where the map isolates population density from the variable non-Hispanic Whites. Furthermore, all covariates used for making maps have been converted to standard normal scores. That is, each covariate is identically independently distributed with mean zero and variance of one for each covariate. Furthermore, we can observe a notable difference in figure 3.9, which is unadjusted versus figure 3.10, which has been adjusted for population density. Due to the redundancy of producing an adjusted map for each covariate, figure 3.10 will be our only illustration.

Fig. 3.11 Full Map with Population Density Isolated

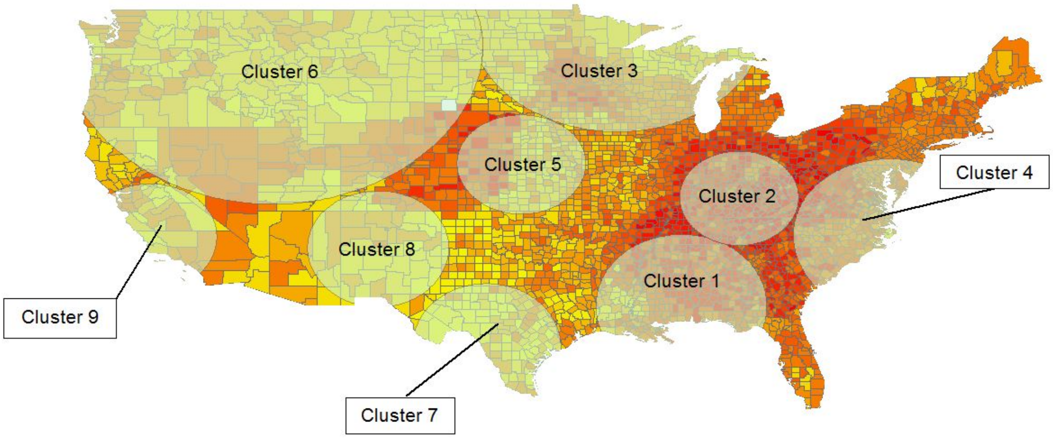


Figure 3.11 illustrates all covariates as a response variable, where the population density has been isolated from each covariate. This implies that population density does not play a role when covariates are clustered together. Again, clusters are ranked by a hierarchy, where cluster 1 is the most likely cluster and cluster n is the n^{th} most likely cluster. Each cluster identified in figure 3.11 is explained in the table 3.8 and 3.9. That is, strictly looking at cluster 1, cluster 1 contains a high rate or above average amount of NATA, pm 2.5, African Americans, Females, insufficient sleep, mentally distressed, and adult obesity. Furthermore, this map only groups covariates together and is not used as a predictor for a given political party.

Table 3.8 Multivariate Clusters, Part I

Cluster	Data Set	Number of Counties	Mean Inside
1	NATA	310	1.28
	PM 2.5	310	0.25
	African American	310	1.37
	Female	310	0.56
	Insufficient Sleep	310	1.03
	Mentally Distressed	310	1.06
	Adult Obesity	310	1.01
2	NATA	310	0.22
	PM 2.5	310	1.44
	Female	310	0.16
	Insufficient Sleep	310	0.79
	Mentally Distressed	310	0.76
	White	310	0.88
	Adult Obesity	310	0.42
3	Catholic	309	0.83
	Excessive Drinking	309	1.25
	White	309	0.65
	Some College	309	0.73
4	NATA	295	0.41
	PM 2.5	295	0.36
	African American	295	0.93
	Female	295	0.61
	Insufficient Sleep	295	0.52
	Mentally Distressed	295	0.061
	Obesity	295	0.011
	Some College	295	0.036

Table 3.9 Multivariate Clusters, Part II

Cluster	Data Set	Number of Counties	Mean Inside
5	Catholic	184	0.42
	Excessive Drinking	184	0.34
	White	184	0.61
	Obesity	184	0.17
	Some college	184	0.75
6	Catholic	309	0.19
	Excessive Drinking	309	0.41
	White	309	0.05
	Some College	309	0.41
7	Catholic	109	0.91
	NATA	109	0.086
8	Catholic	58	1.22
	Mental Distress	58	0.27
9	Catholic	30	1.24
	NATA	30	0.45
	African American	30	0.082
	Female	30	0.46
	Some College	30	0.28

Due to the covariates being normalized, the mean inside for tables 3.8 and 3.9 indicate how many standard deviations a cluster or data set is above the mean compared to the rest of the contiguous counties in the US.

3.3 Factor Analysis

Factor analysis is useful when there are many variables used in a model. Factor analysis allows one to collapse many variables into a few interpretable underlying factors [10]. The objective of factor analysis is to find independent latent variables, that is, to find variables that are not directly observed.

When running a factor analysis, several assumptions must first be met. These assumptions include random errors have a mean of zero, the common factors are standard normal with a mean of zero and variance of one. Furthermore, the common factor is uncorrelated with one another, specific factors are uncorrelated with one another, and specific factors are uncorrelated with common factors. Developing the principal component method **S** is the sample variance-covariance matrix written as

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X_i} - \bar{x})(\mathbf{X_i} - \bar{x})'$$

where

$$\mathbf{X_i'} = [X_{i1}, X_{i2}, \ldots, X_{ip}]$$

We will have p eigenvalues of **S**: $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ and p eigenvectors of **S**: $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p$, for the variance–covariance matrix.

Since the variables or rather common factors need to be standard normal, we convert all our variables to normal quantiles using Blom's method. From here we derive the rotated factor patter as illustrated in table 3.10. The values contained in table 3.10 are called factor loadings. Factor loading show the relationship of each variable to the underlying factor. Furthermore, factor loadings can be interpreted like standardized regression coefficients [10], that is, we can think of factor loadings as a correlation coefficient. The estimator for the factor loadings is given by

$$\hat{l}_{ij} = \hat{e}_{ji}\sqrt{\hat{\lambda}_j}$$

Table 3.10 Rotated Factor Pattern

	Factor 1	Factor 2	Factor 3
Mentally Distressed	0.79912	0.26646	-0.08804
Insufficient Sleep	0.69072	0.48696	-0.19435
Obese	0.60841	0.04869	-0.02578
Catholic	-0.61633	-0.13093	-0.14477
Some College	-0.72668	0.16729	0.19048
Excessive Drinking	-0.76423	-0.11659	-0.02296
Population Density 2014	-0.13553	0.80814	-0.04733
NATA	0.33841	0.67978	-0.31466
African American	0.30924	0.62766	-0.49766
Female	0.09739	0.55010	0.04403
Non-Hispanic White	-0.10650	-0.23596	0.76880
Particulate Mater 2.5	0.28750	0.31526	0.40244

After running a factor analysis on our twelve variables, three factors are yielded. Factor 1 contains mentally distressed, insufficient sleep, obese adults, Catholics, come college, and excessive drinking. We can further identify that Catholics, some college, and excessive drinking have a high negative correlation with mentally distressed, insufficient sleep, and obese. Factor 2 contains insufficient sleep, population density for 2014, NATA, African Americans, and Female. Finally, factor 3 contains non-Hispanic Whites, particulate matter 2.5, and African Americans, where African Americans are negatively correlated with non-Hispanic Whites and PM_{25}.

Now that three factors have been established, the factor scores need to be extracted from each factor for every variable. A factor score indicates an existence on a hidden factor. Factor scores are derived using the ordinary least squares method. The vector of common factors, $\hat{\mathbf{f}}_i$, is found by minimizing the sum of the squared residuals [9]:

$$\sum_{j=1}^{P} \varepsilon_{ij}^2 = (\mathbf{Y_i} - \mu - \mathbf{Lf_i})'(\mathbf{Y_i} - \mu - \mathbf{Lf_i})$$

where

$$\mathbf{Y_i} = \mu + \mathbf{Lf_i} + \varepsilon_i$$

and

$$\mathbf{f_i} = \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}}\hat{\mathbf{e}}_1'(\mathbf{Y_i} - \bar{y}) \\ \\ \frac{1}{\sqrt{\hat{\lambda}_2}}\hat{\mathbf{e}}_2'(\mathbf{Y_i} - \bar{y}) \\ \vdots \\ \\ \frac{1}{\sqrt{\hat{\lambda}_m}}\hat{\mathbf{e}}_m'(\mathbf{Y_i} - \bar{y}) \end{bmatrix}, \mathbf{L} = \left[\sqrt{\lambda_1}\mathbf{e_1}, \sqrt{\lambda_2}\mathbf{e_2}, \cdots, \sqrt{\lambda_m}\mathbf{e_m},\right]$$

The factor scores are then normalized using Blom's method once again. The resulting normal quantiles are then used in SaTScan running the normal model.

Fig. 3.12 Factor 1 Map Adjusted for Population Density

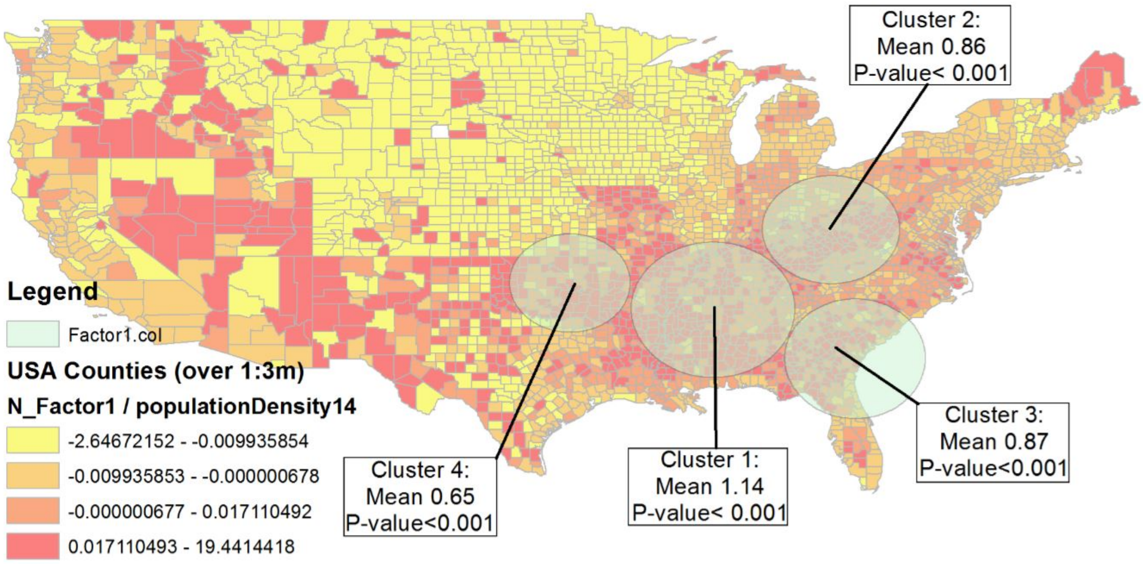


Table 3.11 Factor 1 Adjusted for Population Density

Cluster	Mean Inside	Number of Counties	P-value
1	1.14	310	<0.001
2	0.86	292	<0.001
3	0.87	185	<0.001
4	0.65	115	<0.001

Running a spatial scan for clusters on factor 1, four clusters were identified. These clusters indicated that they are between 1.14 and 0.65 standard deviations above the mean. Furthermore, due to the factor loadings given in Table 3.10, we can identify that the counties analyzed for factor 1 suffer from mental distress, insufficient sleep and adult obesity. We can further identify these counties as having a low percentage of a catholic population, and low percentage of post-secondary education, as well as the counties do not drink excessively.

Fig. 3.13 Factor 2 Map Adjusted for Population Density

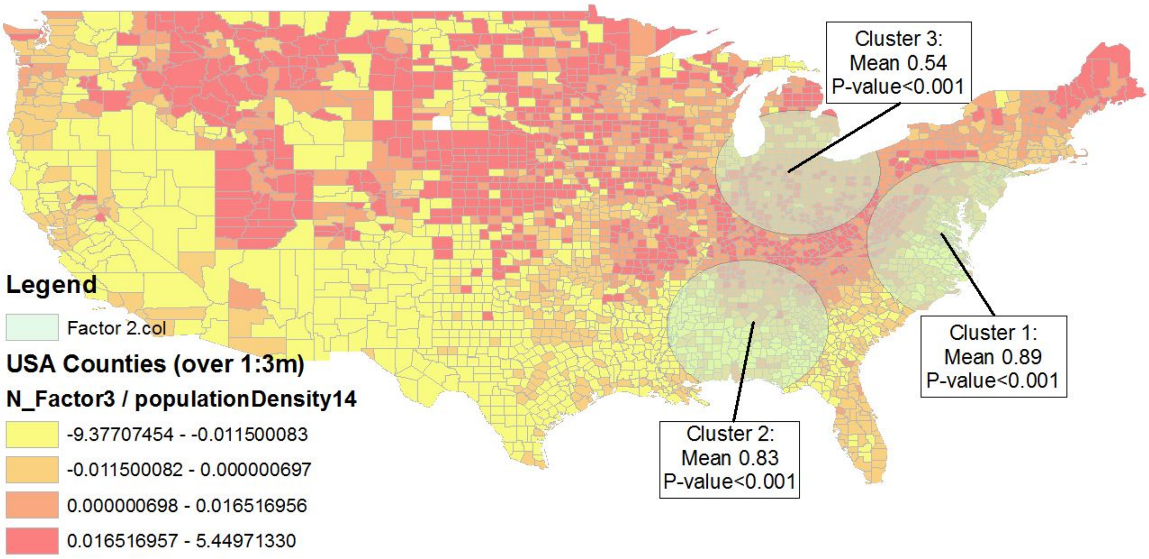


Table 3.12 Factor 2 Adjusted for Population Density

Cluster	Mean Inside	Number of Counties	P-value
1	0.89	308	<0.001
2	0.83	309	<0.001
3	0.54	292	<0.001

The spatial scan for factor 2 identifies three cluster, where these clusters are between 0.89 and 0.54 standard deviations above the mean compared to the reset of the contiguous counties in the US. Based off the factor loadings from Table 3.10, we can identify these counties of having a high population density, high levels of carcinogenic air quality, a high percentage of African American and female population compared to other contiguous counties, and suffer from insufficient sleep.

Fig. 3.14 Factor 3 Map Adjusted for Population Density

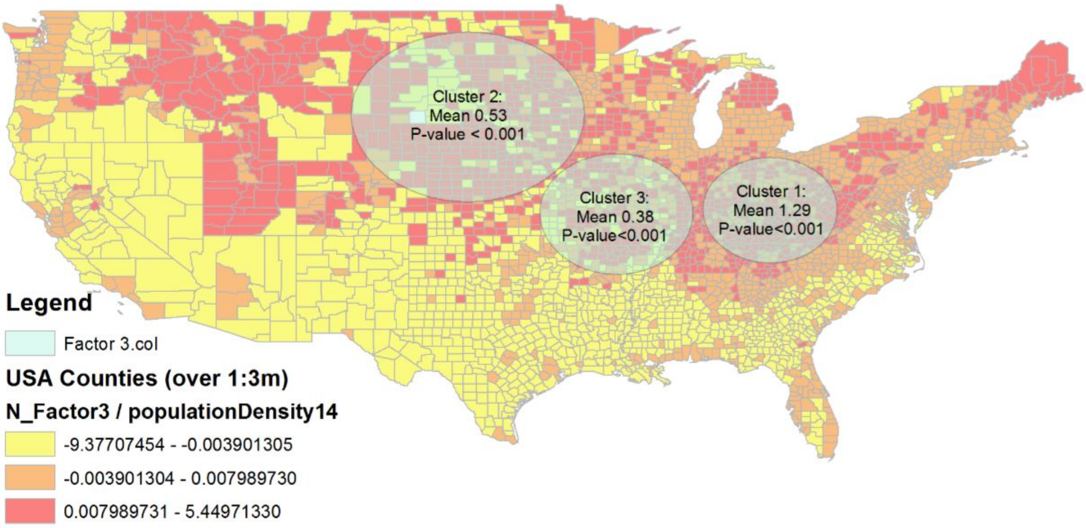


Table 3.13 Factor 3 Adjusted for Population Density

Cluster	Mean Inside	Number of Counties	P-value
1	1.29	310	<0.001
2	0.53	308	<0.001
3	0.38	274	<0.001

Using ArcGIS and SaTScan, three clusters were identified for Factor 3. These clusters ranged from 1.29 to 0.38 standard deviations above the mean compared to the rest of the contiguous counties in the US. These clusters identify a low percentage of African American population compared to the rest of the US, a high percentage of Non-Hispanic Whites compared to the rest of the US, and a high rate of particulate matter 2.5 compared to the rest of the contiguous counties in the US.

3.4 Comparison of Means

Using only a select number of covariates first identified in logistic regression, we compare the means of both political parties. That is, we take the averages of each covariates rate for each political party and take the respective mean for said covariate. Furthermore, each variable was broken up into two groups based off color. Group 0 was given color 0, i.e. democratic party, group 1 was given color 1, i.e. republican party, where the groups represent the treatment and the rates for the covariates are the response values. Then, a series of one-way ANOVAs were run. Each treatment was identified as significant this can be observed in **Appendix B**. Thus, each mean for each variable is different, or rather significant.

Table 3.14 Comparison of Means

Variable	Party	Average Rate	95% Confidence Interval
Obesity	GOP	31.33666	(31.14486, 31.52847)
	DEM	28.93595	(28.48935, 29.38255)
Drinking	GOP	16.43697	(16.29150, 16.58243)
	DEM	17.02624	(16.68753, 17.36495)
College	GOP	55.4005	(54.8993, 55.9017)
	DEM	60.8743	(59.7074, 62.0413)
Stress	GOP	11.14790	(11.05769, 11.23812)
	DEM	11.74050	(11.53044, 11.95056)
Sleep	GOP	32.68590	(32.50949, 32.86231)
	DEM	34.84112	(34.43037, 35.25187)
AA	GOP	6.7003	(6.1184, 7.2822)
	DEM	21.4854	(20.1306, 22.8402)
White	GOP	81.5525	(80.8062, 82.2987)
	DEM	55.0784	(53.3409, 56.8160)
Female	GOP	49.83834	(49.74247, 49.93421)
	DEM	50.59511	(50.37189, 50.81833)

Table 3.14 illustrates the average rate for each covariate for each part, where the average rate is given a 95% confidence interval. We can further observe for every variable listed, we can compare the confidence interval for each political party and see that there are no overlapping confidence intervals. This goes together with the multiple one-way ANOVAs, that is, all the means are considered significantly different from their respective counterpart.

Chapter 4

Conclusion

4.1 Summary

We can conclude that counties that voted republican, rather for Donald Trump, can be described as predominantly having a higher rate of non-Hispanic white people, a lower rate of African Americans, and have a higher rate of obesity among adults. Furthermore, these counties can be continued to be classified as not as educated, a lower rate of excessive drinking, and do not suffer from mental distress. However, counties that voted democratic, that is, for Hillary Clinton, can be described as having a higher rate of African Americans, a lower rate of non-Hispanic white people, and a higher rate of education. Also, these counties drink in excess, have a lower rate of obesity, and suffer from a higher rate of mental distress. Furthermore, we can conclude that the means of each variable are significantly different between political parties. Therefore, the biggest set of attributers for counties that for voted Donald Trump in order of most significant are counties that are mostly white, have very few African Americans, and are obese.

4.2 Suggestions for Further Study

The population used throughout this study pertained to the 2014 population. Utilizing data for the estimated population counts of 2016 may improve upon modeling adequacy. Furthermore, data used from the EPA, both NATA and PM 2.5 was from the year 2011. As population grows, industry grows. We might expect the carcinogenic values of NATA to increase. Therefore, a study conducted with more recent NATA values would be beneficial and may also increase model adequacy. When running an ANOVA on African Americans, female, and non-Hispanic Whites the residuals were not normally distributed. Perhaps a transformation of the data for each of these variables is needed. Furthermore, it may prove useful to determine how age groups voted, e.g. do certain age groups tend to vote more for one party versus the other?

References

[1] (n.d.). *County Health Rankings Home Page.* N.p. Web. Jan. 4, 2017.

[2] ArcGIIS (n.d.). *ArcGIS online Home Page.* N.p. Web. Nov. 4, 2016.

[3] EPA (n.d.). *Environmental Protection Agency Home Page.* N.p. Web. Oct. 30, 2016.

[4] Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., and Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 u.s. presidential election. *Analyses of Social Issues ond Public Policy*, 9(1):241–253.

[5] Hsieh, F. Y., Bloch, N., and Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *STATISTICS IN MEDICINE*, 17(17):1623—1634.

[6] Kulldorff, M. (2015). Satscan user guide. (version 9.4).

[7] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis.* John Wiley & Sons, INC., 5 edition.

[8] PennState (Web. Apr. 2, 2017a). Lesson 12: Factor analysis. https://onlinecourses.science.psu.edu/stat504/node/163.

[9] PennState (Web. Apr. 2, 2017b). Receiver operating characteristic curve (roc). https://onlinecourses.science.psu.edu/stat505/node/74.

[10] Rahn, M. (Web. Apr. 2, 2017). Factor analysis: A short introduction, part 1. http://www.theanalysisfactor.com/factor-analysis-1-introduction/.

[11] SAS (n.d.). *SAS Home Page.* N.p. Web. Nov. 4, 2016.

[12] Silver, N. (2016). A user's guide to fivethirtyeight's 2016 general election forecast. *FiveThirtyEight*.

[13] Sinha, P., Sharma, A., and Singh, H. (2012). Prediction for the 2012 united states presidential election using multiple regression model. *The Journal of Prediction Markets*, 2(6):77–97.

Appendix A

A.1 List of Covariates

Table A.1 Variable Labels and Names

Label	Variable Name	Used In Model
Rate_Asmoking	Adult Smoking	No
Rate_Aobes	Adult Obesity	Yes
Rate_excesDrinking	Excessive Drinking	Yes
Rate_Hsgrad	High School Graduate	No
Rate_someCollg	Some College	Yes
Rate_unemp	Unemployment	No
Rate_VioCrime	Violent Crimes	No
Rate_menDistrss	Mentally Distressed	Yes
rate_insuffSleep	Insufficient Sleep	Yes
rate_unins	Uninsured	No
Median_HouseInc	Median Household Income	No
rate_ovr65	Over 65 years of Age	No
rate_AA	African American	Yes
rate_nonHisWhite	Non-Hispanic Whites	Yes
rate_female	Female	Yes
PovertyPercent	Poverty	No
CathRate	Catholic Rate	Yes
PM_25	Particulate Matter	Yes
NATA_RAW	NATA	Yes
populationDensity14	Population Density for 2014	Yes

A.2 List of Response Variables

Table A.2 Response Labels and Names

Label	Variable Name
Color	Political Party Identifier
Trump	Percent of votes Trump Received per County
Votes_GOP_2016	The count of votes Trump Received per county

Appendix B

B.1 ANOVAs

Table B.1 Obesity

Source	DF	SS	MS	F Value	P-value
Color	1	2355.092716	2355.09272	122.69	<0.0001
Error	3106	59623.56762	19.19625		
Total	3107	61978.66033			

Table B.2 Excessive Drinking

Source	DF	SS	MS	F Value	P-value
Color	1	141.89320	141.89320	12.85	0.0003
Error	3106	34295.03101	11.04154		
Total	3107	34436.92421			

Table B.3 Some College

Source	DF	SS	MS	F Value	P-value
Color	1	12243.6711	12243.6711	93.42	<0.0001
Error	3106	407085.7743	131.0643		
Total	3107	419329.4454			

Table B.4 Mentally Distressed

Source	DF	SS	MS	F Value	P-value
Color	1	143.49594	143.49594	33.79	<0.0001
Error	3106	13190.83475	4.24689		
Total	3107	13334.33069			

Table B.5 Insufficient Sleep

Source	DF	SS	MS	F Value	P-value
Color	1	1898.05990	1898.05990	116.89	<0.0001
Error	3106	50435.97007	16.23824		
Total	3107	52334.02997			

Table B.6 African American

Source	DF	SS	MS	F Value	P-value
Color	1	89325.8724	89325.8724	505.61	<0.0001
Error	3106	548733.5045	176.6689		
Total	3107	638059.3769			

Table B.7 Non-Hispanic Whites

Source	DF	SS	MS	F Value	P-value
Color	1	286396.907	286396.907	985.61	&lt;.0001
Error	3106	902536.738	290.578		
Total	3107	1188933.645			

Table B.8 Female

Source	DF	SS	MS	F Value	P-value
Color	1	234.02219	234.02219	48.80	<0.0001
Error	3106	14895.14202	4.79560		
Total	3107	15129.16421			