What Characterizes Communities that Support Donald Trump

Henry Zuo

I. Introduction

The 2016 Republican Primaries have been nothing but unusual. Donald Trump, a billionaire businessman with no experience holding political office, has become the center of attention in these primaries, both because of his often incendiary remarks and his surprising performance state after state. The advantage of Trump has not been insignificant. He has led the national polling average since October 2015 and never once relinquished his leadership position. Establishment candidates have been beaten by Trump in their home states, despite access to long-time donors, local organizations, and lobby groups of the party political machine. Pundits, including some serious statisticians from fivethirtyeight.com, predicted him to fail time after time, as Trump made controversial remarks about Mexican immigrants, Muslims, women, torture, and a variety of subjects. However, what are normally be considered political gaffes did not hurt Trump's support. Instead, the primary process has seen Trump winning a majority of the states and building a sizeable advantage over his closest competitors.

What characterizes communities that gave rise to support for Donald Trump? If Trump is indeed accomplishing something rarely seen from other Republican candidates before, is there anything that distinguishes his voter base from those of others? These are questions worth answering but remain underexplored. Existing academic literature explaining Trump's success is few and far between. Some media reports used surveys to identify distinguishing traits in Trump voters. A common surveying tool is traditional exit polls, which are conducted in voting sites on voters who just casted their votes. One of such polls was explored by Zitner, who discovered that the demographics and political views of Trump supporters constitute a new coalition that differs from the traditional Republican base (2015). In other research project, Wolfe, Yeip, and Zitner found that Trump supporters were shown to share some characteristics that cut across different camps within the Republican party: they are low-income groups, non-degree holders, against immigration, against free trade, anti-gay-marriage, for gun rights, and non-religious (2015). This conclusion supported the earlier Zitner research by identifying the more specific traits of Trump voters. Not all research projects support these hypotheses. Zhen pointed out that Trump supporters might have been demonized by such statistics and the main stream media at large, that Trump supporters consist of a coalition that incorporates high-income and highly educated groups, that the stereotype of an uneducated, low-income, angry Trump voter was an image conjured up and popularized by the media (2016).

Most existing research largely focuses on sampling a subsection of voters to dissect the composition of Trump supporters. However, the kind of communities surrounding the lives of these Trump supporters deserve more attention. Political support for a candidate does not appear in a vacuum. The political,

economic, cultural, and demographic environment in a community must be able to shape the views of members of the community. This paper distinguishes itself from the overwhelming majority of existing research by focusing on the community rather than the individuals. This paper examines defining traits of communities that give rise to support for Trump. First, this paper will collect and organize demographic data that characterize counties in 25 states that finished their Republican Primaries, as well as voting results in these Republican Primaries. Then, this paper will explore the relationship between demographic traits and support for Trump by finding the best model that predicts support for Trump. Lastly, this paper will discuss the significance of its research and future applications.

II. Methods

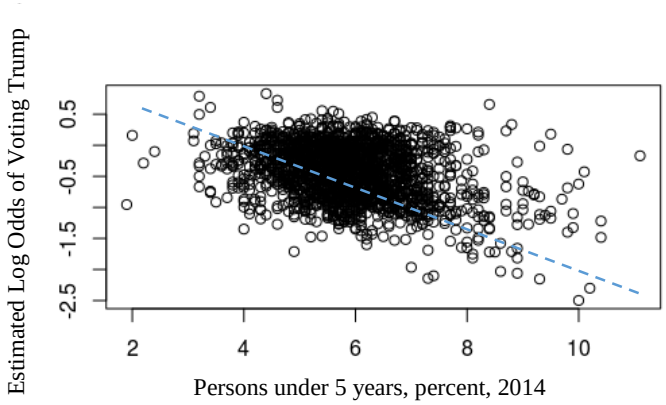
The Republican Primaries are still ongoing, but detailed data for 25 of the 50 states were available at the time when this research project started. These results were collated by contributors on Kaggle.com and organized in .csv files for people to research on (see the reference section for source of data). This paper will use data of primary voting results of these 25 states, and demographic information about these 25 states. The full list of all 25 states can be found in Table A in appendix, and highlighted in Figure 1 below:

Figure 1. Primary States Included by Dataset, as Highlighted

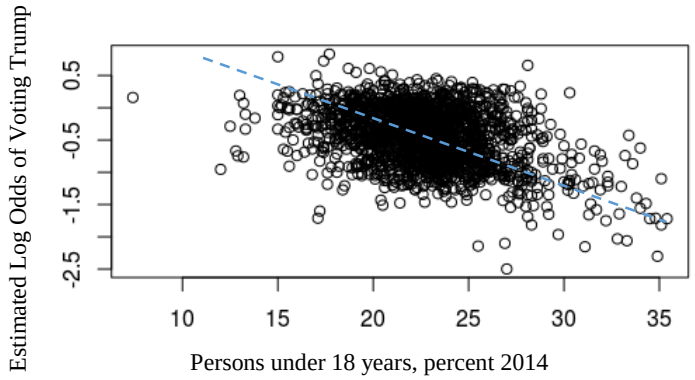
Ideal data for this research would include the demographic variables of the living environment of every voter. Data with such a high level of detail are impossible to collect. Realistically, demographic variables at the county level are available, and this paper can investigate the research question if it treats voters

from the same county as experiencing the same environment. With this assumption, the paper gathers information from two datasets from Kaggle.com, and produces a combined dataset. The combined dataset has 43 variables and 1881 observations. It contains demographic information of every county that voted in the 25 states that this paper investigates. It also contains the number of people who did and did not vote for Trump in the 2016 primary. Demographic variables that are not meaningful for the purpose of understanding the characteristics of communities that support Trump are removed. For details on how the final combined dataset is created, or a table of all predictors available, see appendix.

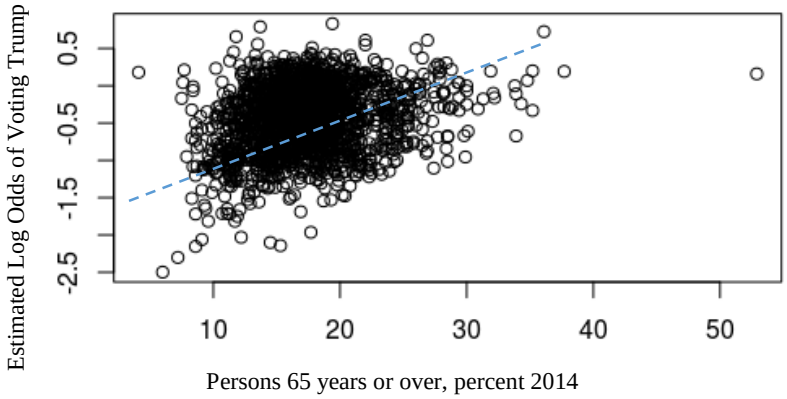
This paper uses a logistic model to examine the relationship between support for Trump and demographic variables at the county level. In a logistic model, support for Trump is represented by what is called a "log odds", with the following mathematical form: $\log\left(\frac{p}{1-p}\right)$. P in this formula represents the probability that a voter would vote for Trump. $\frac{p}{1-p}$, which is the probability of voting for Trump over the probability of not voting for Trump, is called odds. $\log\left(\frac{p}{1-p}\right)$ is odds that are logged. After this paper sets the log odds as the proxy for Trump support, the paper explores the dataset. The paper plots the log odds of voting for Trump against every demographic variable of all counties. Based on observing the patterns of the plots, the paper discovers six demographic variables that appears to change with the log odds of voting for Trump. The six plots are shown in Figure 2 below, where the dashed line is the trend in each plot observed that this paper observed:



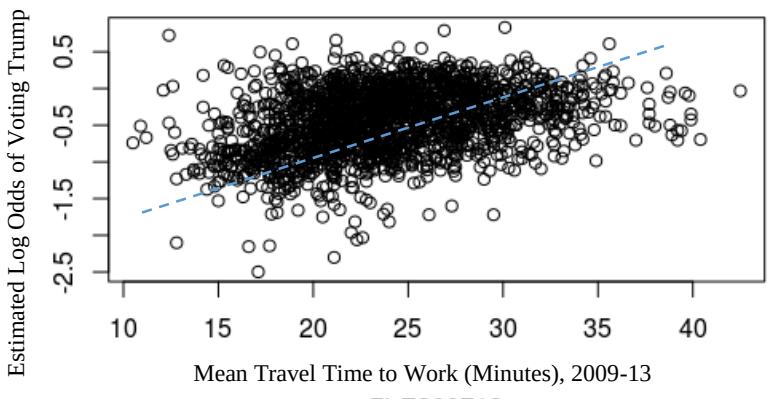
We observe that as the percentage of people under 5 years old increases, the log odds of the fraction of Trump votes generally decreases.



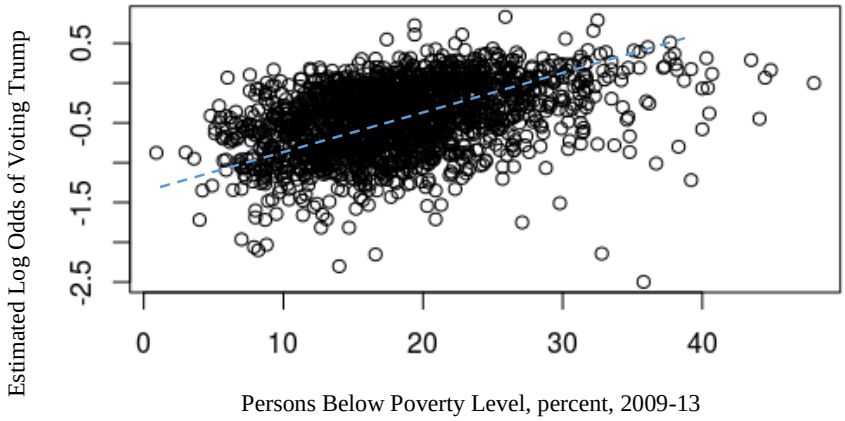
We observe that as the percentage of people under 18 years old increases, the log odds of the fraction of Trump votes generally decreases.



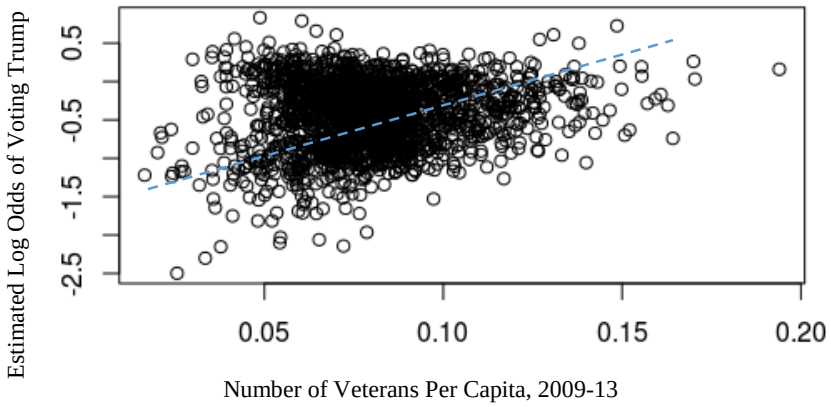
We observe that as the percentage of people over 65 years old increases, the log odds of the fraction of Trump votes generally increases.



We observe that as Mean travel time to work (minutes) increases, the log odds of the fraction of Trump votes generally increases.



We observe that as percentage of people below the poverty line increases, the log odds of the fraction of Trump votes generally increases.



We observe that as number of veterans per capita increases, the log odds of the fraction of Trump votes generally increases.

Figure 2. Variables Shown to Correlate Strongly with Trump Support from Data Explorations

After data explorations, this paper can model the log odds of a voter voting for Trump based on available demographic variables. We will include a random effect for each state, because voters in counties within the same states are likely to be similar. The mathematical form of the model is shown below:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u_{State\,1} + u_{State\,2} + \cdots + \varepsilon \qquad (1)$$

The left side of the model is the log odds that represents support for Trump, as explained previously on page 4 of this paper. On the right side, the *x*s are demographic variables that characterize communities that witnessed strong support for Trump. The *β*s represent the coefficient of these demographic variables. The *u*s are random intercepts for the state that each county is in. This is needed because voting results within the same state tend to correlate with each other, as voters in the same state have similar media exposure, experience similar campaign coverage, experience similar local campaign events, and experience all these campaign efforts at similar time periods. This intrastate correlation tends to affect the results we see for each *β*, which is not desirable. Thus, to better understand the effects of demographic variables, this paper removes the effect of the states by giving every state a random intercept. To make sure the random intercept for states are significant additions, this paper conducts a likelihood ratio test, and confirms that the probability that the random intercept adds no meaningful information is almost zero. The next step this paper takes is picking out the most significant variables out of the all available variables. This is done through something called forward selection. Forward selection adds variables to a

model as long as added variables are meaningful, until the variable to be added does not offer more meaningful information. Every new variable is selected out of all available variables and must create a new model that offers more information than any other model created by other available variables. The criterion used to determine whether a new variable added is meaningful is called Akaike Information Criterion (AIC). When all else is equal, the lower the AIC, the better the model. Frequently, two models with AIC differences smaller than 10 are considered not meaningfully different. However, this standard cannot be applied in this research. In this research, AICs are often in the orders of 5000 to 10,000, because the sample size in this research becomes very large when the logistic regression model treats every individual voter as an independent data point in its calculation. If this paper only excludes a variable when the AIC difference is smaller than 10, all variables will be included. This paper intends to find the most significant and defining characteristics of communities that give rise to support for Trump. Thus, an AIC difference of 3000 is adopted. If the model with the lowest AIC has an AIC that differ from the original AIC by at least 3000, we proceed with the new model. Under this method, this paper finds a final model. The random intercept for the states has a variance of 0.109. All fixed coefficients of the final model are shown in Table 9 below:

Coefficients in Final Model	$\hat{\beta}$	Standard Error	Odds Ratio (exp($\hat{\beta}$))
Intercept	0.207100	0.066260	1.230106
Bachelor's degree or higher, percent of persons age 25+, 2009-2013	-0.031350	0.000111	0.969136
Hispanic-owned firms, percent, 2007	-0.007455	0.000073	0.992573
Persons under 5 years, percent, 2014	-0.080080	0.000682	0.923043
Asian alone, percent, 2014	0.025960	0.000287	1.026300
Per capita money income in past 12 months (2013 dollars), 2009-2013	0.000019	0.000000	1.000019

Table 9. All Fixed Variables in Final Model

How reliable are these results? This paper proceeds with a series of tests to make sure the model fits the observed data well. Information offers by all tests overall showed that the final model performed well. See section c of the appendix for more details.

III. Results and Discussion

The final model gives us several pieces of information. There are three variables that correlate negatively with support for Trump. The estimated odds of support for Trump change by a factor of 0.9691 for every unit increase in percentage of “degree-holders”, people with a bachelor’s degree or higher, in the county. The estimated odds of support for Trump change by a factor of 0.992573 for every unit increase of percentage of Hispanic-owned companies. The estimated odds of support for Trump change by a factor of 0.923043 for every unit increase in percentage of population of five-year-olds. Two variables correlate

positively with support for Trump. The estimated odds of support for Trump change by a factor of 1.026300 for every unit increase in percentage of Asians. The estimated odds of support for Trump change by a factor of 1.000019 for every dollar increase in per capita money income.

The final model confirmed only one out of the six discoveries made in data explorations. This showed that support for Trump is highly complex, with many demographic variables at play. It is possible that many demographic variables are highly correlated, so the observed relationships in the graphs in data explorations might not be valid once variations in other variables are considered in the mode. It is also possible that all of the six variables explain changes in Trump support. The model selection process for the final model is highly selective, demanding an AIC difference of at least 3000. While this high standard makes sure the significance of variables that end up in the final model, other less significant but valid variables are left out, in the tradeoff between fit and parsimony of model selection.

Over all, the paper discovered several big trends. First, the paper confirmed the racial component behind Trump support, something that has been popularly reported by the media. Counties that are diverse, in this case counties characterized by a higher percentage of Asian population, tend to vote for Trump. It is possible that when Republican primary voters get more exposure to minority, they are more likely to vote for Trump. Counties that have minorities controlling more social resources, in this case counties with a higher percentage of Hispanic-owned companies, tend to vote against Trump. This could suggest that local financial and political resources in these counties are deployed against Trump. This could also mean that when minorities gain higher socioeconomic status (as Hispanic business owners in this case), Trump's call to deport Hispanic immigrants becomes less appealing to Republican primary voters. Second, the widely perpetuated media narrative that Trump supporters are generally less educated segments of the public who never received a college degree is confirmed. This paper discovers that when more members of a county holds a college degree or above, voters are less likely to support Trump. Third, economic woes are, contrary to common belief, not necessarily a source of Trump's support. This paper discovered that higher per capita income in 2013 correlates with higher chance to vote Trump. Forth, support for Trump certainly varies from state to state. This paper discovers that there is a 0.106 variance across states in log likelihoods of probability to vote Trump.

The paper used a very strict criterion to choose variables in its final model. An AIC difference of 3000 is an uncommonly harsh standard. This is primarily because the logistic regression model made the actual population available to the model very large. Larger population generally has more variability. It becomes much less likely to reject the null hypothesis, the hypothesis that there is no relationship between a predictor and the response variable. The adoption of an unusually high standard is meant to prevent an overly large model. After all, the objective of the research is to understand characteristics of communities

that give rise to support for Trump. The final model is not meant to make accurate predictions in every future primary county. It is more important to be concise and identify the most significant demographic variables than get any variable that could be relevant. Alternatively, this paper could treat percentage of Trump votes as the probability that a given voter in a given county would vote for Trump, and get the log odds of voting for Trump through the probability at each county. This could a model setup for future research.

IV. Reference

Hamner, B. (2016, March 28). 2016 US Election | Kaggle. Retrieved April 29, 2016, from https://www.kaggle.com/benhamner/2016-us-election

How Trump Happened. (n.d.). Retrieved April 29, 2016, from http://graphics.wsj.com/elections/2016/how-trump-happened/?mod=e2fb

P., & Zitner, A. (n.d.). Donald Trump Forges New Blue-Collar Coalition Among Republicans. Retrieved April 29, 2016, from http://www.wsj.com/articles/donald-trump-forges-new-blue-collar-coalition-among-republicans-1449272326

Zhen, X. (2016, April 20). My Roundtable with Trump Supporters in New York. Retrieved April 29, 2016, from http://zhuanlan.zhihu.com/p/20729724?refer=iamelection

V. Appendix

a. Data Manipulation

The paper gathers information from two datasets from Kaggle.com. The first dataset, *primary_results*, breaks down the voting results of each candidate-Republicans and Democrats alike-by county in the 2016 party primaries. Two variables in this dataset describe the voting results. The first one is the number of votes received by a candidate. The second one is the fraction of the total votes casted in the county received by a candidate. An sample entry is shown below:

state	state_abbreviation	county	fips	party	candidate	votes	fraction_votes
Alabama	AL	Autauga	1001	Republican	Donald Trump	5387	0.445

Table 1. Sample Entry of *primary_results*

The second main dataset used by this paper, *county_facts*, describes the demographic information of every county in America. It has 51 variables, each of which describe one attribute of the demographics of the counties. These attributes range from retail sales per capita to percentage of college-degree holders. A sample entry is shown below:

fips	area_name	state_abbreviation	PST045214	PST040210	PST120214	POP010210	AGE135214	…
1001	Autauga County	AL	55395	54571	1.5	54571	6	…

Table 2. Sample Entry of *county_facts* (46 columns omitted)

All predictors in the *county_facts* dataset are listed Table 3 in the appendix section.

Three tasks of data manipulation are needed. First, information from the two separate datasets must be combined before this paper can carry out statistical analysis. This paper combined the two datasets through a variable called the Federal Information Processing Standard (FIPS) county code, which is a five-digit code that uniquely identifies counties and county equivalents in the United States. FIPS is available for every county in both datasets. The FIPS code connects the voting results in each county in *primary_results* with demographic information of each county in *county_facts*.

Second, Dataset *county_facts* offers the demographic information this paper needs, but dataset *primary_results* does not offer immediately usable information about support for Trump: it only offers the fraction votes received by Trump, as shown in Table 4 below:

fips	area_name	state_abbreviation	PST045214	PST040210	…	votes	fraction_votes
1001	Autauga County	AL	55395	54571	…	5387	0.445

Table 4. Sample Entry of Combined Dataset

However, fraction of votes is not an ideal measure of support for Trump for the purpose of building a quantitative model. This is because if this paper builds a model that predicts percentage of support for Trump using linear regression, the model could predict a support percentage that goes beyond 100%, which is impossible in reality, but possible in a linear model. A more ideal model is a logistic model, with a response variable of whether an average voter in a county supports Trump, represented by zero and one. This model is consistent with reality, since its prediction of the probability that an average voter would support Trump is between 0% to 100%. How does this paper get to the response variable of the ideal model? This paper determined the number of votes that did not go to Trump in every county based on the number of Trump votes and fraction of Trump votes. Then, in the process of model building, this paper will treat each individual vote as an independent data entry. Thus, while a sample entry of the combined dataset is what was shown in Table 4, a sample data entry of the actual model will be different, as shown in Table 5 below:

fips	area_name	state_abbreviation	PST045214	PST040210	…	Whether a Voter Voted for Trump
1001	Autauga County	AL	55395	54571	…	1

Table 5. Sample Entry of Combined Dataset as Used by Model

Third, not all variable in the dataset is a meaningful predictor. For instance, variables such as Total Food Services Sales and Total Private Nonfarm Employment are only meaningful if they are viewed in the context of the counties' population sizes. This is because the values of these variables are naturally greater when total population of a county is larger, and vice versa. Comparisons between absolute values between these variables are not meaningful. Thus, this paper converted such variables into per-capita numbers. The list of these variables are listed as follows:

column_name	description
AFN120207	Accommodation and food services sales, 2007 ($1,000)
BPS030214	Building permits, 2014
BZA010213	Private nonfarm establishments, 2013
BZA110213	Private nonfarm employment, 2013
MAN450207	Manufacturers shipments, 2007 ($1,000)
NES010213	Nonemployer establishments, 2013
VET605213	Veterans, 2009-2013
WTN220207	Merchant wholesaler sales, 2007 ($1,000)

Table 6. Variables Converted into Per-Capita Numbers

Lastly, some predictors are not quite meaningful for the purpose of characterizing a community to understand how it gives rise to political views. To improve the efficiency of this research project, they are removed from the dataset. These variables are listed as follows:

PST045214	Population, 2014 estimate
PST040210	Population, 2010 (April 1) estimates base
POP010210	Population, 2010
HSG010214	Housing units, 2014
HSD410213	Households, 2009-2013
RTN130207	Retail sales, 2007 ($1,000)
SBO001207	Total number of firms, 2007
HSG495213	Median value of owner-occupied housing units, 2009-2013

Table 7. Variables Removed

b. Tables and Figures

Full list of 25 states	AL, AR, AZ, FL, GA, IA, ID, IL, KY, LA, MA, MI, MO, MS, NC, NH, NV, OH, OK, SC, TN, TX, UT, VA, VT

Table 7. List of All Primary States Investigated

Column Name	Description
PST120214	Population, percent change - April 1, 2010 to July 1, 2014
AGE135214	Persons under 5 years, percent, 2014
AGE295214	Persons under 18 years, percent, 2014
AGE775214	Persons 65 years and over, percent, 2014
SEX255214	Female persons, percent, 2014
RHI125214	White alone, percent, 2014
RHI225214	Black or African American alone, percent, 2014
RHI325214	American Indian and Alaska Native alone, percent, 2014
RHI425214	Asian alone, percent, 2014
RHI525214	Native Hawaiian and Other Pacific Islander alone, percent, 2014
RHI625214	Two or More Races, percent, 2014
LFE305213	Mean travel time to work (minutes), workers age 16+, 2009-2013
HSG445213	Homeownership rate, 2009-2013
HSG096213	Housing units in multi-unit structures, percent, 2009-2013
SBO015207	Women-owned firms, percent, 2007
RTN131207	Retail sales per capita, 2007
RHI725214	Hispanic or Latino, percent, 2014
RHI825214	White alone, not Hispanic or Latino, percent, 2014
POP715213	Living in same house 1 year & over, percent, 2009-2013
POP645213	Foreign born persons, percent, 2009-2013
POP815213	Language other than English spoken at home, pct age 5+, 2009-2013
EDU635213	High school graduate or higher, percent of persons age 25+, 2009-2013
EDU685213	Bachelor's degree or higher, percent of persons age 25+, 2009-2013
LND110210	Land area in square miles, 2010
POP060210	Population per square mile, 2010
PVY020213	Persons below poverty level, percent, 2009-2013
BZA115213	Private nonfarm employment, percent change, 2012-2013

SBO115207	American Indian- and Alaska Native-owned firms, percent, 2007
SBO315207	Black-owned firms, percent, 2007
SBO215207	Asian-owned firms, percent, 2007
SBO515207	Native Hawaiian- and Other Pacific Islander-owned firms, percent, 2007
SBO415207	Hispanic-owned firms, percent, 2007
HSD310213	Persons per household, 2009-2013
INC910213	Per capita money income in past 12 months (2013 dollars), 2009-2013
INC110213	Median household income, 2009-2013

Table 8. All Predictors in *county_facts*

c. Tests for Final Model

This paper conducts a lack-of-fit test based on the chi-squared distribution of the deviance of the final model. With degrees of freedom at 1872, the model has a deviance of 159658.4, which puts the p-value of this test at close to zero. This means that if the null hypothesis is true (that the model does not exhibit lack of fit), the probability of getting the deviance at least this large is very low. This shows that the deviance is too large. There is strong evidence that the final model exhibits lack of fit. The paper then plots the deviance residuals of the model against the predicted log odds of probability of someone voting for Trump. The plot is shown in Figure 3 below:

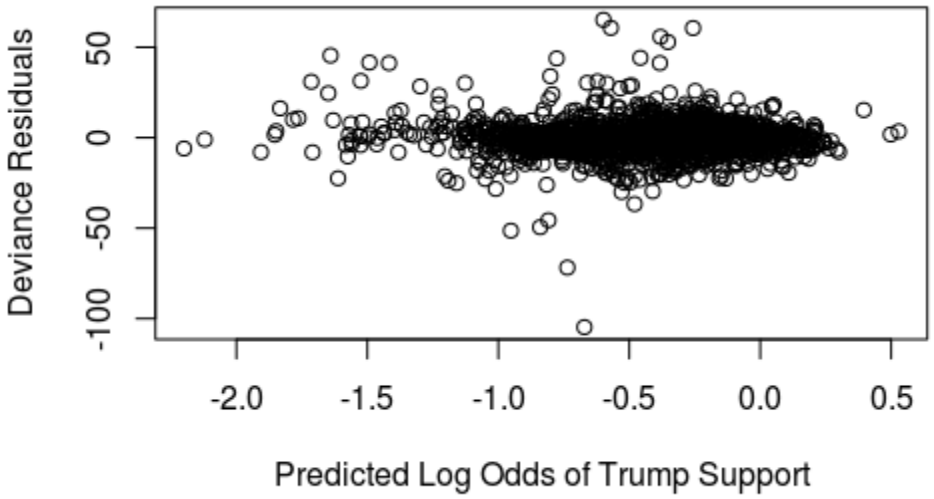


Figure 3. Deviance Residual Against Predicted Values Plot

The plot shows very little pattern. Deviance residuals area generally randomly distributed regardless of the values of the predicted log odds of Trump support. This shows that the model fits the data quite consistently. The paper then checks for outliers in the data using a half-normal plot. This plots the absolute residuals against the quantiles of a normal distribution. The plot is shown in Figure 4. below:

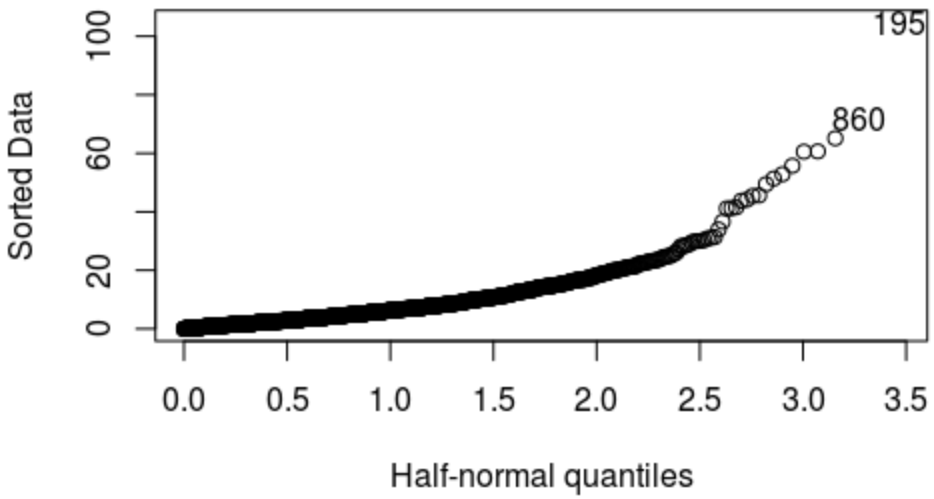


Figure 4. Half-normal Plot of Final Model

The half-normal plot does not show any outliers. Overall, the model fits the data generally well. For the purpose of understanding the most significant characteristics of communities that give rise to support for Trump, the model performs well.