Correlation between the Presence of Big Box Retailers and a County's Political Leanings

Pouya Mohammadi

Introduction

In recent election cycles, election analysts such as Dave Wasserman have pointed to an interesting metric and its being highly correlated with the results of United States (US) presidential elections. This metric is whether the county contains a Whole Foods or a Cracker Barrel (Wasserman, 2020). This has lead to a flurry of analyses about how the presence of different big box retailers is correlated to that area's political leanings. NBC News, Time Magazine, and The New York Times have all recently reported on similar trends with varying reatilers, and on Twitter after the 2020 election, there was speculation that the presence of a Trader Joe's in a county was becoming a valid indicator for the direction in which a county would vote in the election ("Broke: Joe", 2021).

These trends are important as computational politics becomes of ever-increasing importance in modern political campaigns. In elections that are decided by the slimmest of margins, such as the US presidential election of 2000 that was decided by less than 600 votes, every informational advantage that a campaign has can be of use and drastically influence the politics and future of the United States and the world (Glass, 2018). By understanding the relationships between the presence or absence of certain big box retailers in a county and the political preferences of that county, campaigns may be able to draw insights into the ways that communities' shopping habits correspond with their politics and more effectively and efficiently communicate with potential voters. In addition to providing insight into the politics of communities, this method of using the presence of big box retailers in a county to infer information about politics is computationally efficient and far less expensive than big data methods that rely on storing and analyzing information on an individual basis.

Our case study aims to analyze the presence of a new set of big box retailers in US counties and understand the ways that the presence of these counties corresponds with that county's political leanings. In particular, the stores that we will be focusing on are Trader Joe's, Cinemark, Cabela's, and Nordstrom. These stores were chosen in part because of the fact they were easily accessible data sources, but mostly because of our belief that they will capture information about the different voting blocks in the US electorate incredibly well. Trader Joe's targets singles, couples, and small families living in larger cities or the areas surrounding large cities (Watson, 2014). Cabela's tends to target individuals who tend to be white, conservative males. These include hunters, fishermen, military personnel, and law enforcement (Martin, 2013). Nordstrom's target market is high-end shoppers and the middle class (Bhogaraju, 2015). We suspect that this corresponds with suburban voters who are an important voting bloc. Finally, Cinemark's target market is "midsized markets or suburbs of major cities" (Team, 2020). We believe that the midsized market corresponds to an important voting bloc that is not encapsulated in the target market of the other retailers.

Aside from Trader Joe's, all of these stores have yet to be the subject of a rigorous analysis that relates their presence to political leanings. Trader Joe's was included because we believe that it corresponds to a different voting block than the other stores that we chose and the fact that the body of work on the political leanings of "Trader Joe's counties" is new and does not include a methodology similar to ours. Our null hypothesis is that we do not expect there to be a relation between the presence of any of the stores and the political leanings of the counties in which these retailers are located.

Data

We use publicly available data online that can be downloaded for the list of sites in the Data subsection of our References. These include open-source government databases as well as public location listings by private companies. The 2020 presidential election results that we use to measure a county's political leanings are borrowed from the GitHub user @tonmcg, who developed the datasets from information provided by The Guardian, townhall.com, Fox News, Politico, and the New York Times (Tonmcg, 2020). All of the government provided databases are estimations of the current values based on the most US Census which occurred in 2010 and trends in the data since that time. Finally, the location listings are either provided by the stores themselves or reputable third-party sources, such as Fandango. We use the 'geopy' Python library to map these locations to their corresponding geographical county. Any locations whose counties cannot be found by the Python library are hand-encoded. Because Alaska uses electoral districts instead of counties in their voting procedures, we will not consider Alaska in this analysis and remove all data from Alaska in our dataset prior to analysis. While this solution is not ideal, it is consistent with previous literature and prudent since we do not feel comfortable using two different region distinctions in the calculation of our dependent variable (Gomez et al., 2007).

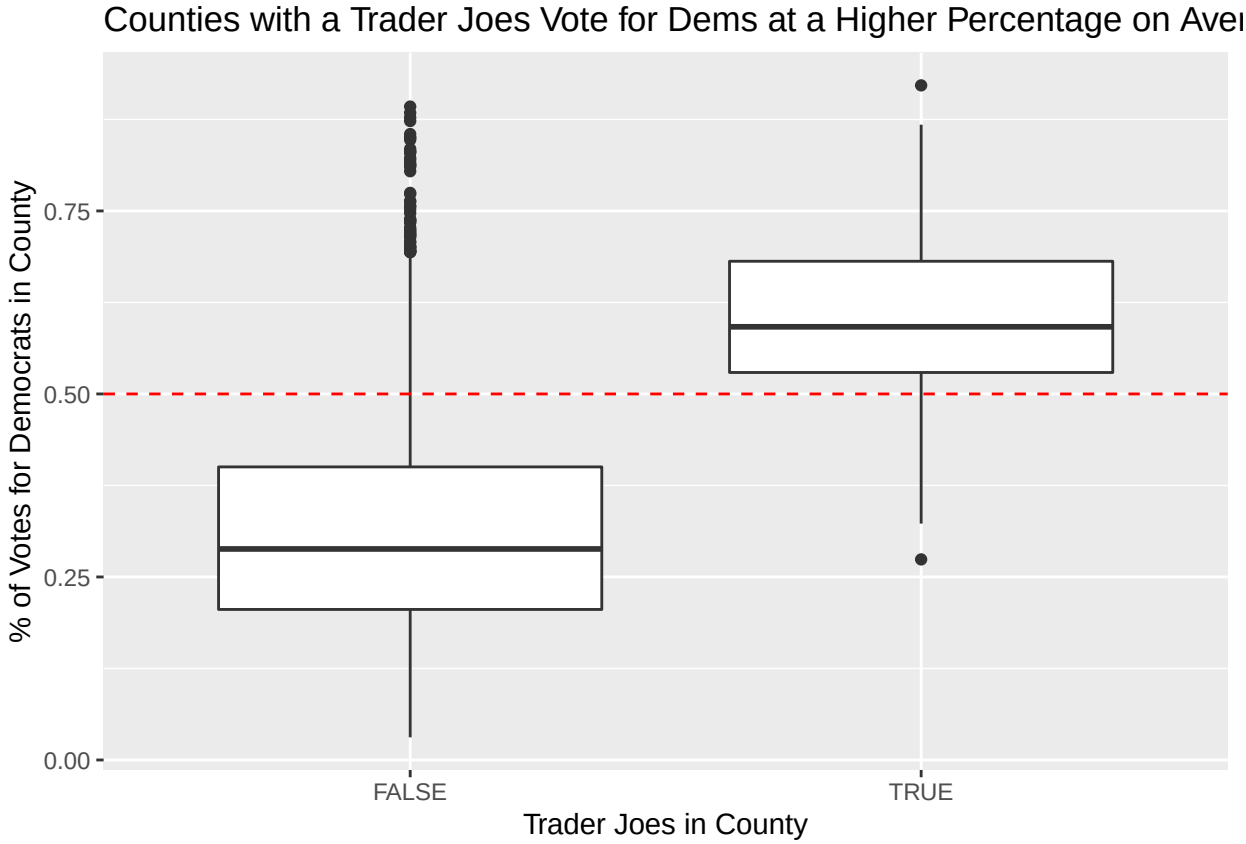
We will use the percentage of votes for the Democratic party in the 2020 presidential election results from a county as our dependent variable, which will be a numeric variable between 0 and 1, to measure the political leanings of a county. For our independent variables, we will have four binary variables that each correspond to whether a county contains one of the four retailers that are the focus of this analysis. Each county will have a 1 for the respective variable if the county contains the retailer and a 0 if it does not contain the retailer. In addition to these variables, we will include indpendent variables that are usually included in analyses of a county's political leanings. These variables are borrowed from Kahane, and they include population totals by race, population totals by gender, education levels, poverty rates, the median income of the county, unemployment rates, and whether the county is urban or rural (Kahane, 2020). We exclude certain variables that we did not have access to, such as religion in our analysis.

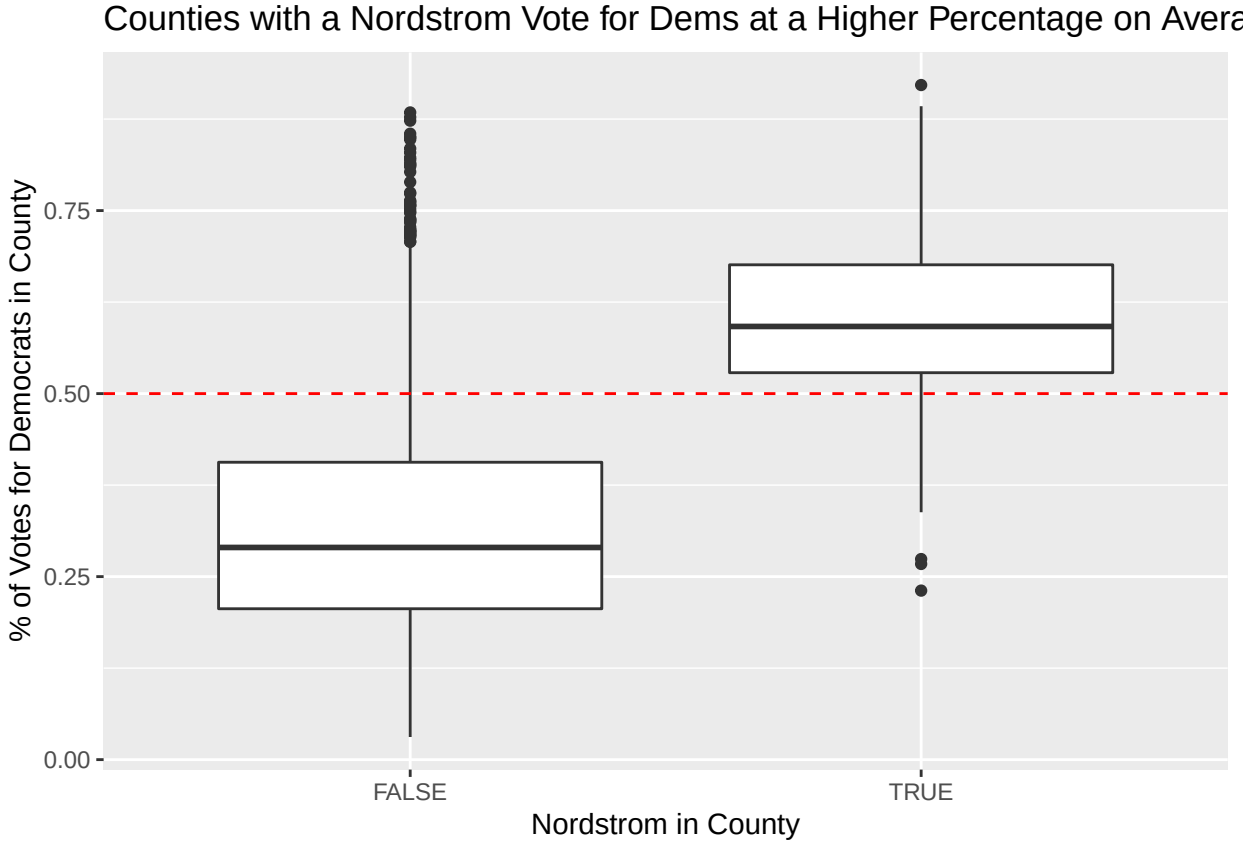
Existing Literature

As mentioned previously, plenty of news networks have completed analyses about the ways that big box retailers' locations correspond to political leanings in a county. In particular, NBC News, Time Magazine, and The New York Times have conduced analyses on this phenomenon. The New York Times in their analysis, found that, without controlling for other factors, the presence of a Trader Joe's corresponded with better results for Democrats. The analysis discovered that Democrats won areas that were within five miles of a Trader Joe's by 33 points. Time magazine discovered a similar trend in that Democrats won districts with a Trader Joe's by 30 points in 2014 US congressional races. A different study by Aaron Lee that employs a random forest machine learning model that uses 20 big box retailers discovered that counties that contain a Trader Joe's lean democratic with a feature importance of about 0.11 (Lee, 2020).

Despite all of the attention to these trends, there does not seem to a published work on the topic. We expand on these past analyses by analyzing a new set of big box retailers and their presence's relation to political leanings in a county. We also control for other variables that are customarily used in predicting a county's political leaning, which has yet to be done in a scientific analysis. We gather this list of variables from Kahane, and we explore whether the presence of these big box retailers will provide more significant relationships to a county's political leaning than the standard covariates.

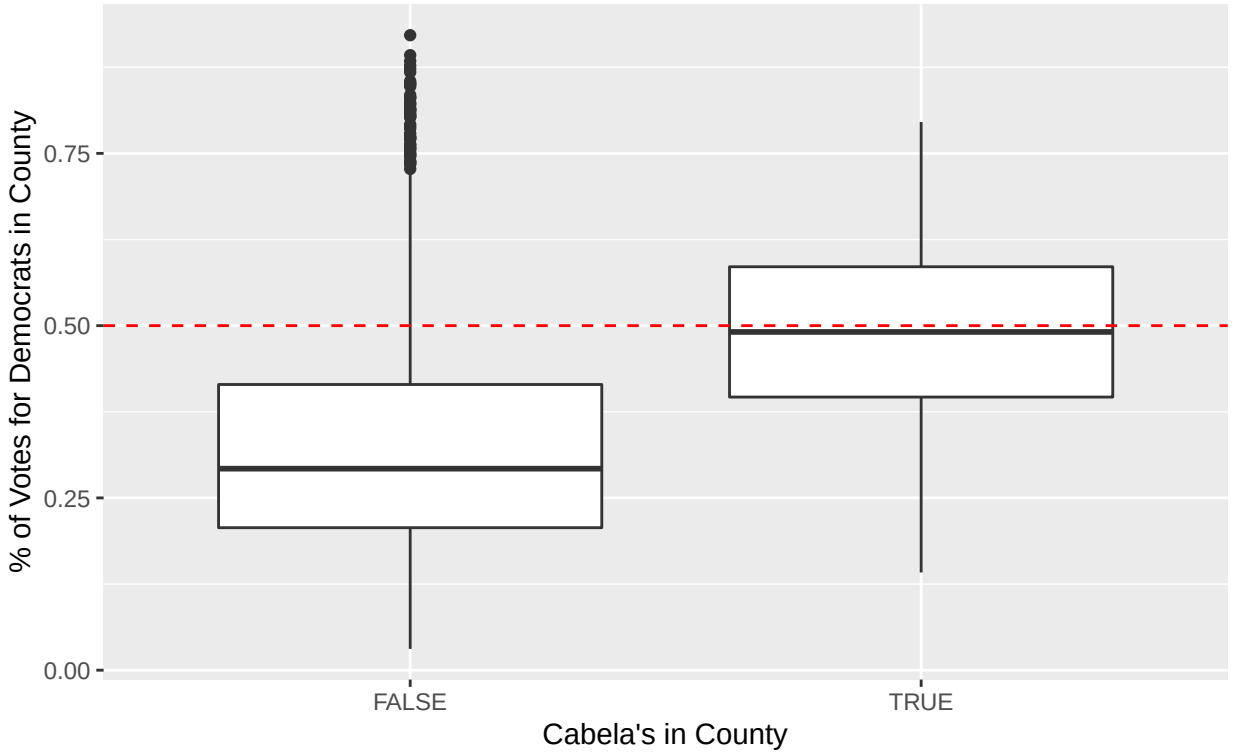
EDA


Counties with a Trader Joes Vote for Dems at a Higher Percentage on Aver
% of Votes for Democrats in County
0.75
0.50
0.25
0.00
FALSE
TRUE
Trader Joes in County


Counties with a Nordstrom Vote for Dems at a Higher Percentage on Avera
% of Votes for Democrats in County
0.75
0.50
0.25
0.00
FALSE
TRUE
Nordstrom in County

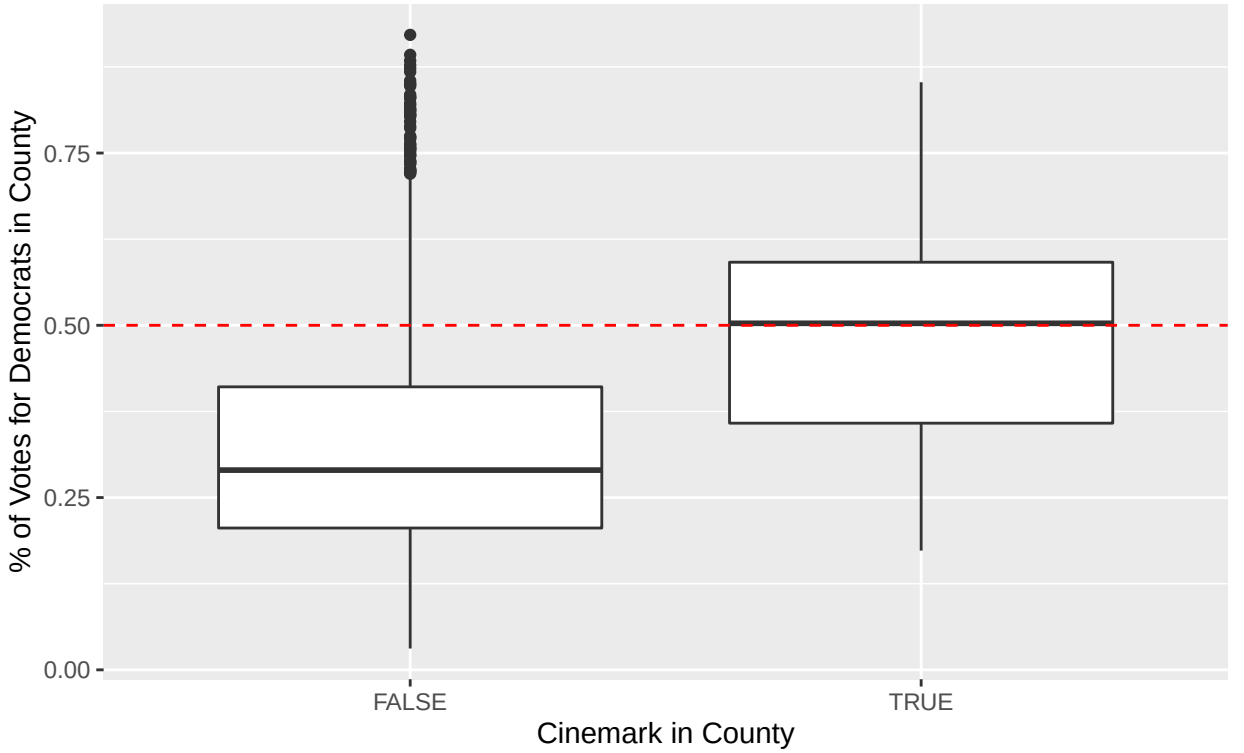
Counties with a Cabela's Vote for Dems at a Higher Percentage on Averag

However, Democrats did not receive a majority of the votes in counties with a Caebla's



Counties with a Cinemark Vote for Dems at a Higher Percentage on Averag

Democrats received 50% of the Vote in Counties with a Cinemark on average



These graphs show that regardless of the retailer, it appears as if Democrats performed better in counties that contained the retailer more often Republicans. Although, this trend appears to be greater for Trader Joe's and Nordstrom than Cabela's and Cinemark.

Methodology

Motivation

In order to analyze whether the presence of big box retailers corresponds with a county's political leanings, we will utilize a beta regression model using the percentage of votes in the county cast for Democrats during the 2020 president election as our dependent variable. We chose this dependent variable since the percentage of votes that a party receives is the best measure of a county's political leanings in the United States, given that the United States is a largely two-party political system. By using the percentage of votes for Democrats, we can capture information that we would not be able to if we simply used a binary indicator for which party won the county. For instance, a county that Democrats won by 0.2 percentage points would be regarded equally to a county that Democrats won by 50 percentage points, when in reality, one county is a swing district and the other is a staunchly Democratic county. We use the percentage of Democratic votes instead of Republican votes since we believed it would increase interpretability given that our EDA indicated that most counties with one of our four stores will mostly vote for Democrats. However, in our sensitivity analyses in the appendix, we repeat our analysis using the percentage of Republican votes in a county and find that the results are largely the same because of the two party system in the United States.

Our analysis employs a beta regression model. However, other models were considered as well, such as a linear regression model, an OLS model, a poisson model with an offset, and a negative binomial model with an offset. We do not use a linear regression model since our response variable is bounded between 0 and 1, which a linear regression model does not account for. We do not use the OLS model because our dependent variable has values that are close to 0 and as a result, the OLS model could run into the same issue that the linear regression model faces where it does not bound the dependent variable distribution between 0 and 1. Finally, the percentage of votes in a county is a count of the votes for a party divided by the count of total votes in the county, which can be modeled via a model of count data with an offset. In these models, our dependent variable would be the count of votes for a party and the offset would be the total votes in that county. However, we see that a Poisson model is not viable since the mean count of votes for both parties is not equal to the variance of the count of votes for that party, an assumption of the poisson model. We then consider a negative binomial model which assumes that the mean of the dependent count variable is not equal to its variance. However, when performing some EDA, which can be found in the Appendix, we found that the count of votes for either party does not fit a negative binomial distribution despite the fact that it is count data with overdispersion. Instead, the beta regression model assumes that the response variable follows a beta distribution, a distribution of values between 0 and 1. Our response variable does the same and has a single mode, which is similar to a beta distribution again. This is further discussed and visualized in Appendix. In general, the beta distribution is incredibly applicable for proportion data, such as the percentage of votes for a party in a county (Ferrari et al., 2004). Since the only other assumption for a beta regression model is that there is a linear relationship between the predictors and the response variable, we feel confident employing a beta regression model, since the percentage of democratic votes in a county falls between 0 and 1, similar to a beta distribution.

We check the assumptions of the beta regression model in the appendix. The beta regression model has the following three assumptions:

1. The response variable, the percentage of votes that Democrats received in a county, follows a beta distribution.
2. There is independence between the observations in our dataset.
3. There is linearity between the predictors and response variable in our model.

Model Formula

education levels, poverty rates, the median income of the county, unemployment rates, and whether the county is urban or rural

$$\begin{aligned} link(\mu_i) =& \beta_0 + \beta_1 * (Trader.Joes.in.County_i = True) + \beta_2 * (Cabelas.in.County_i = True) + \\ & \beta_3 * (Cinemark.in.County_i = True) + \beta_4 * (Nordstrom.in.County_i = True) + \\ & \beta_5 * (White.Population) + \beta_6 * (Black.Population) + \beta_7 * (Hispanic.Population) + \\ & \beta_8 * (Asian.Population) + \beta_9 * (Male.Population) + \beta_{10} * (Female.Population) + \\ & \beta_{11} * (Percent.with.College.Education) + \beta_{12} * (Percent.in.Poverty) + \\ & \beta_{13} * (Percent.Unemployed) + \beta_{14} * (Urban.County = True) \end{aligned}$$

where μ_i is an observation-specific mean for $Percent.Democratic_i$ and we separately estimate the precision parameter for the distribution on the mean to be $\phi_i = precision\_link(\phi_i)$.

Sensitivity Analyses

We will perform a few sensitivity analyses in which we compare the performance of different models. The first model that we will test is replacing the response variable in our original model with the percentage of republican votes in a county in order to see if there is no significant difference between these two models as we hypothesized by assuming that the two-party system would make third-party votes negligible. We will conduct another sensitivity analysis in which we use a ANOVA test with our original model and a model that does not contain any store data in order to see if there is a significant difference in a traditional model of county affiliation and one that includes the store locations.

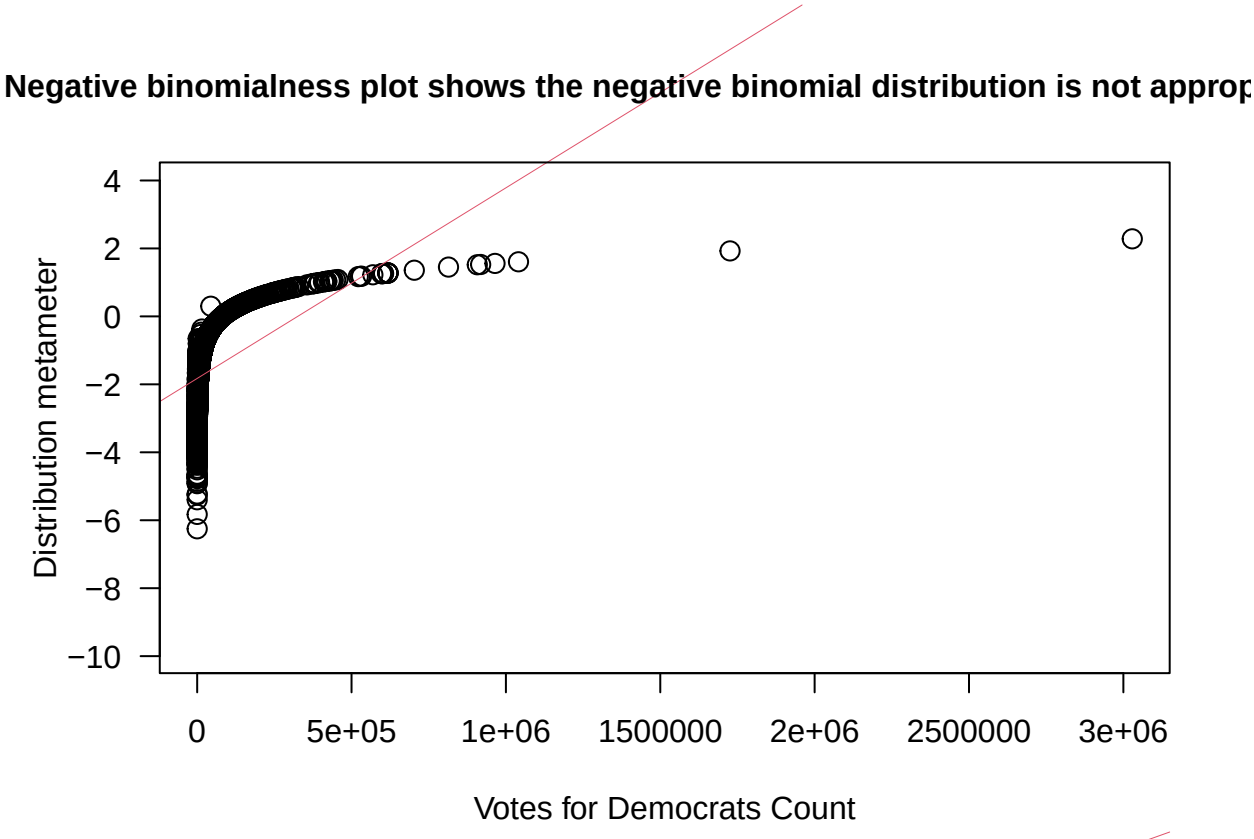
Appendix

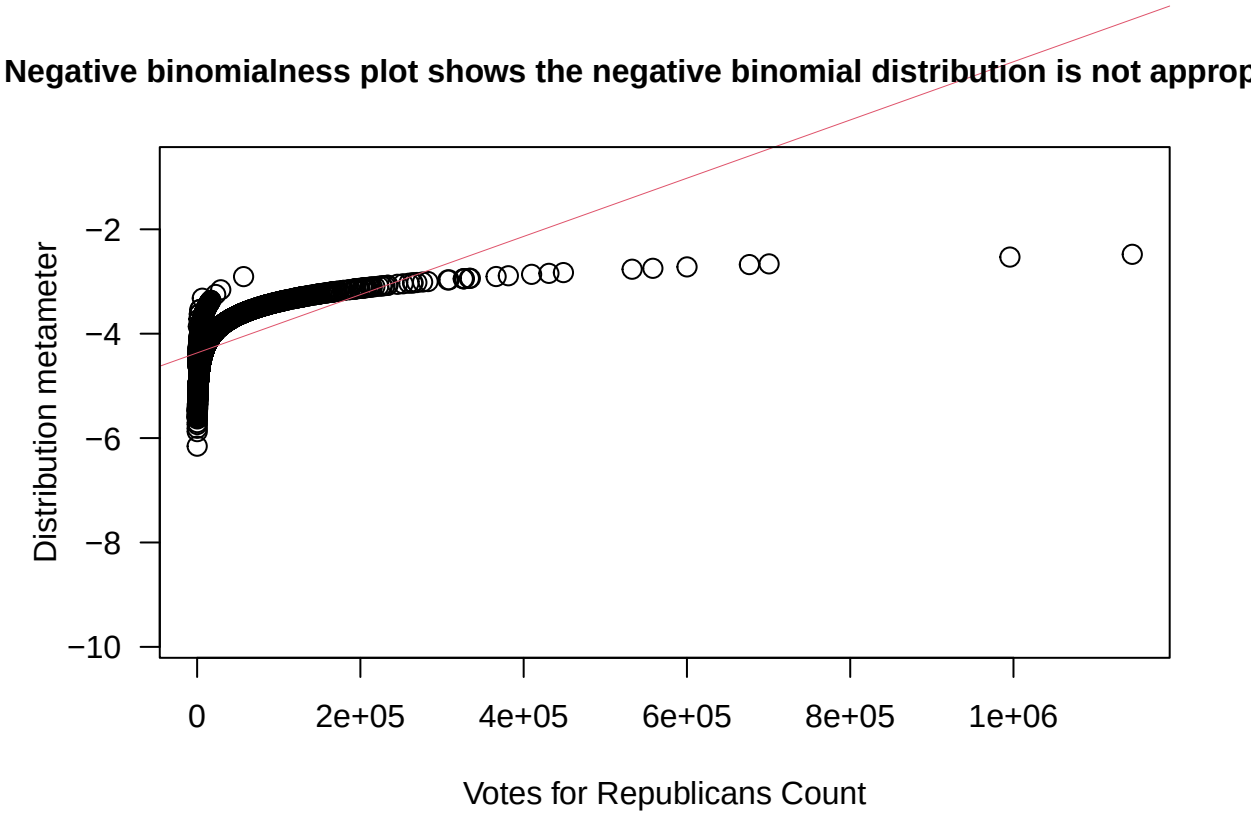
Model Motivation

We see from the plots below that the mean and variance of the votes for each party are not equivalent indicating that while the data is count data, it does not follow the poisson distribution and a poisson regression cannot be used here.

	Democratic Vote Count	Republican Vote Count
mean	26064.1	23784.8
variance	9509979014.6	2953216930.8
ratio	364868.4	124164.0

From the plots below, we see that neither the votes for Democrats in a county nor the votes for Republicans in a county follows a negative binomial distribution. If a negative binomial model was appropriate, there would be a linear relationship between the distribution metameter and the count of votes for a respective party. However, we see that there is a logarithmic relationship between these variables in our plots below.


Negative binomialness plot shows the negative binomial distribution is not approp
Distribution metameter
4
2
0
−2
−4
−6
−8
−10
0
5e+05
1e+06
1500000
2e+06
2500000
3e+06
Votes for Democrats Count

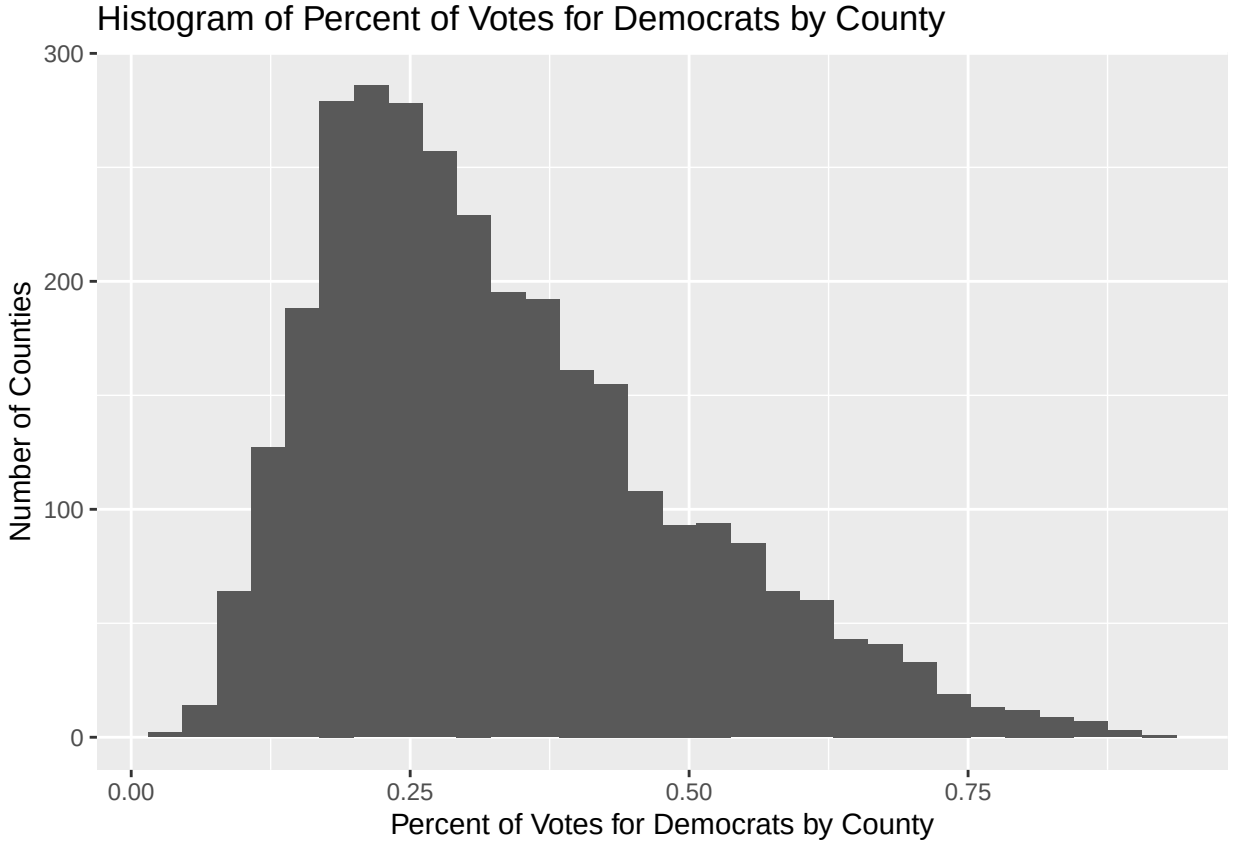

Negative binomialness plot shows the negative binomial distribution is not approp
Distribution metameter
−2
−4
−6
−8
−10
0
2e+05
4e+05
6e+05
8e+05
1e+06
Votes for Republicans Count

Model Assumptions and Diagnostics

We will check the three assumptions for the beta regression model below. The three assumptions are as follows:

1. The response variable, the percentage of votes that Democrats received in a county, follows a beta distribution.
2. There is independence between the observations in our dataset.
3. There is linearity between the predictors and response variable in our model.

Below, we see that both the percentage of votes that Democrats received in a county have a range between 0 and 1, with a single mode. Since these characteristics are the same as the characteristics of a beta distribution, we feel confident modeling the percentage of votes in a county for Democrats as a beta distribution.



Next, we analyze the assumption that our observations are independent of each other. While there are some concerns here regarding individuals in different counties of a similar makeup being exposed to the same media and neighboring counties being in similar communities that affect the politics of the county, we feel confident that these effects are negligible and that we include the proper covariates in our model to account for any confounding variables. Plotting the residuals below, we see that there is a random scatter of residuals and this helps us feel more confident that there is independence between our observations.

We will next investigate whether the linearity assumption is met for our model by plotting the fitted means for our observed data and the residuals of our data. Seeing that there is a random scatter around the y-axis, we feel confident that the linearity assumption is met.

Next, we will check for potentially influential points in our model. We plot both the leverage and the Cook's distance values of our model, and we see that we have a few observations that have a high leverage. However, looking at the Cook's distance for these values, we see that there are no points with a high Cook's distance. As a result, we feel confident that there are no influential points in our model.

References

Bhogaraju, Sirisha. “Nordstrom’s Target Customers.” Market Realist. Market Realist, February 18, 2015. https://marketrealist.com/2015/02/nordstroms-target-customers/.

Ferrari, Silvia, and Francisco Cribari-Neto. “Beta Regression for Modelling Rates and Proportions.” Journal of Applied Statistics 31, no. 7 (2004): 799–815. https://doi.org/10.1080/0266476042000214501.

Glass, A. (2018, December 12). Bush declared Electoral victor Over Gore, Dec. 12, 2000. Retrieved March 03, 2021, from https://www.politico.com/story/2018/12/12/scotus-declares-bush-electoral-victor-dec-12-2000-1054202

Gomez, Brad T., Thomas G. Hansford, and George A. Krause. “The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections.” The Journal of Politics 69, no. 3 (2007): 649–63. https://doi.org/10.1111/j.1468-2508.2007.00565.x.

Kahane, L. H. (2020). Determinants of County-Level voting patterns in the 2012 and 2016 presidential elections. Applied Economics, 52(33), 3574-3587. doi:10.1080/00036846.2020.1713985

Lee, A. (2020, September 29). Trader Joe’s Democrats and Walmart Republicans. Retrieved March 04, 2021, from https://towardsdatascience.com/are-you-a-trader-joes-democrat-or-a-walmart-republican-a7b156131435

M. (2021, February 16). Broke: Joe Biden did so well in counties with a Trader Joe’s because the audience for Trader Joe’s is composed of favorable Democratic demographics Woke: Joe Biden did so well in counties with a Trader Joe’s because voters thought he was Trader Joe. Retrieved March 03, 2021, from https://twitter.com/maxtmcc/status/1361504297477890050

Martin, Cindy. “Cabela’s.” CE Martin, November 10, 2013. http://cemartin.weebly.com/uploads/1/8/9/1/18911943/cabelas_marketing_assignment.docx#:~:text=Cabela’s%20target%20markets%20include%20avid,law%20enforcement

Team, MBA Skool. “Cinemark SWOT Analysis: Top Cinemark Competitors, STP & USP: Detailed SWOT Analysis of Brands.” MBA Skool-Study.Learn.Share. MBA Skool, April 12, 2020. https://www.mbaskool.com/brandguide/media-and-entertainment/15016-cinemark.html.

Wasserman, D. (2020, December 08). Fact: Biden won the presidency Winning 85% of counties with a Whole foods and 32% of counties with a Cracker barrel - the widest gap ever. Retrieved March 03, 2021, from https://twitter.com/Redistrict/status/1336342894630858755

Watson, Elaine. “Quirky, Cult-like, Aspirational, but Affordable: The Rise and Rise of Trader Joe’s.” foodnavigator. William Reed Business Media Ltd., April 15, 2014. https://www.foodnavigator-usa.com/Article/2014/04/15/Quirky-cult-like-aspirational-affordable-The-rise-of-Trader-Joes#:~:text=Trader%20Joe’s%20targets%20singles%2C%20couples%2C%20and%20small%20families%20%E2%80%8B&text=less

Data

Trader-Joes-Stores.pdf. (2020). Retrieved March 03, 2021, from https://www.traderjoes.com/pdf/Trader-Joes-Stores.pdf

Nordstrom Store Addresses. (2020, June 1). Retrieved March 03, 2021, from http://nordstromsupplier.com/Content/sc_manual/Store_Address_List.pdf

Cinemark Movie Theater Locations. (2021). Retrieved March 04, 2021, from https://www.fandango.com/movie-theaters/cinemark

All Cabela’s Locations: Sporting goods & outdoor stores. (2021). Retrieved March 04, 2021, from https://stores.cabelas.com/

Small Area Income and Poverty Estimates (SAIPE). (n.d.). Retrieved March 04, 2021, from https://www.census.gov/data-tools/demo/saipe/#/?map_geoSelector=mhi_c&s_measures=mhi_snc&s_year=2019

Economic Research Service - Download data. (2021, February 24). Retrieved March 04, 2021, from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

Tonmcg. (2020). Us_county_level_election_results_08-20. Retrieved March 03, 2021, from https://github.com/tonmcg/US_County_Level_Election_Results_08-20