ADS Final Project: Examining Variables to Provisional Ballot Reject

Github repository with all code:

https://github.com/nbc270/Provisional-Ballot-Investigation

Christine Biddlecombe (cb4221)
Nathan Caplan (nbc270)
Ursula Kaczmarek (uak211)

Abstract

In this report, we examine the effects of various measures on provisional ballot rejection rates. Specifically, we explore the inference that removing the requirement to vote at designated precincts lowers the provisional ballot rejection rate. As part of this exploration, we isolate the effect of Hurricane Sandy, a storm which significantly affected the midterm and presidential elections in November 2012. Our findings suggests there is a correlation between removing polling location requirements and lower provisional ballot rejection rates. We can consider provisional ballot rejection as an indication some eligible voters are unaware of the location of their polling locations, and that voter communication from the board of elections is an area of needed improvement.

Introduction

Provisional ballots allow individuals who do not appear on the voter roll to cast votes that are only counted once the board of elections confirms eligibility.[1] Compared to other states, New York has high usage and rejection rates for this type of ballot. Rejection of a provisional ballot suggests that ineligible voters may be attempting to vote illegally or that eligible voters are attempting to vote legitimately, but are at the wrong polling place. In the first case, a high provisional ballot rejection rate suggests provisional ballots act as a security measure that enhances the integrity of the vote

[1] N.Y. Elec. Law § 8-302.

count. In the second case, a high rejection rate signals that a board of elections has either not disseminated clear communication to voters or has mismanaged the voter roll.

Our inquiry focuses on determining the factors that contribute to New York City's high provisional ballot rejection rate. We examine the conduct of general elections in all five boroughs for the years 2012, 2014, and 2016. In the realm of election administration, we pay particular attention to the ratio of poll workers to eligible voters, and, given the correlation between provisional ballot use and age established in the literature, the proportion of new election-year registrants in New York City under the age of 25.[2]

In addition to examining features consistent with election administration and the electorate, we study the effects Hurricane Sandy had on voting across several federally-declared disaster states: Delaware, Maryland, New Jersey, New York, Ohio, Pennsylvania, Virginia, and the District of Columbia. To facilitate voting among displaced residents, the governors of New York and New Jersey issued executive orders allowing voting by provisional ballot anywhere in the state, while the remaining states did not make similar allowances. The executive orders act as a natural control on the use of provisional ballots outside of designated precincts. and we explore the inference that removing the requirement to vote at designated precincts lowers the provisional ballot rejection rate. Specifically, we examine counties across these states that were certainly affected by the storm (such as the five NYC boroughs, where damages from the storm reached $19 billion[3]) and counties such as Erie and Genesee counties in upstate New York, far removed from the weather effects of Sandy.

[2] Shaw, D., & Hutchins, V. (2013, June 1). Report on Provisional Ballots and American Elections. In *CALTECH/MIT VOTING TECHNOLOGY PROJECT*.

[3] Martin Z. Braun and Freeman Klopott, Bloomberg Seeks $9.8 Billion Aid for NYC Sandy Storm Losses (November 26, 2012.) Retrieved December 16, 2018 from https://www.bloomberg.com/news/articles/2012-11-26/bloomberg-seeks-9-8-billion-in-aid-for-nyc-s-sandy-storm-losses

Data Wrangling and Feature Engineering

The U.S. Election Assistance Commission (EAC) administers the Election Administration and Voting Survey (EAVS) following every federal election. Among other subjects, the survey documents election administrators' responses to questions regarding the number of eligible voters, the number of poll workers, the number of provisional ballots cast, and the number of provisional ballots rejected in their jurisdictions.

From the EAVS data, we engineered our target variable by dividing the total number of provisional ballots filed by the number that were rejected. In addition, we standardized the poll worker count by dividing the total number of eligible voters by the number of poll workers.

Data regarding the 2012 elections in the aftermath of Hurricane Sandy came from the this survey data and focused on alternative methods of casting a ballot that could have some effect on the use of provisional ballots, namely no-excuse absentee voting and in-person early voting. Where these options are available, voters could rely on these methods rather than resorting to a provisional ballot on Election Day.

We supplemented EAVS data with additional voter-related data from the New York State voter list, which is made available for academic research through a Freedom of Information Law request to the New York State Board of Elections.[4] The voter roll was the only available source on voter age data, and we created a feature regarding the proportion of election-year registrations that included voters under 25 years of age.

EAVS data were downloaded and munged using the R Tidyverse[5] and readxl[6] packages.

[4] N.Y. Elec. Law § 3-103(5)

[5] Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

[6] Hadley Wickham and Jennifer Bryan (2018). readxl: Read Excel Files. R package version 1.1.0. https://CRAN.R-project.org/package=readxl

Data were complete for years 2014 and 2016. 2012 EAVS data on eligible voter population and poll worker totals was unavailable and was instead derived from the Board of Elections in the City of New York's 2012 Annual Report.[7] Information for Philadelphia poll workers was pulled from the 2013 Election Day Fact Finding Report[8]

Demographic data was retrieved from the American Community Survey, administered by the U.S. Census Bureau. Variables of interest included the percentage of the population identifying as white, the percentage of households that do not speak english well, median income, and population 18 years or older. Values were collected at the NYC borough level for years 2012, 2014, and 2016 1-Year Data Surveys. We chose these variables because of presumed biases among poll workers that can make it more difficult for the voter to participate in the election process. For instance, in 2014, NYC poll workers were instructed to speak English, even if they knew a different language. This prompted backlash from elected officials, who argued that this directive prevented constituents from casting ballots.[9] Further, race, age, and income features are common among most survey methodologies in determining patterns of voting.[10]

Data Issues

For many counties of interest, the data were missing or incorrect. For example, Boston, Massachusetts, reported one provisional ballot submitted in 2012. Surrounding jurisdictions including Brookline, Braintree, and Cambridge reported zero provisional ballots. In other instances, one or two values were missing, most often number of poll workers for a given jurisdiction. Philadelphia, for example, reported no poll workers to EAVS for the 2012 or 2014 elections, requiring us to find the information elsewhere.[11]

[7] New York City Board of Elections, Annual Report 2012 (2012). Retrieved December 18, 2018: http://vote.nyc.ny.us/downloads/pdf/documents/boe/AnnualReports/BOEAnnualReport12.pdf

[8] Nutter Administration, Election Day Fact-Finding Report. (June 18, 2013.) Retrieved December 18, 2018, from https://www.phila.gov/Newsletters/ElectionDayFactFindingReport2013.pdf

[9] Ibid.

[10] File, T. (2017, May 10). Voting in America: A Look at the 2016 Presidential Election. In *Census Blogs*. Retrieved from https://www.census.gov/newsroom/blogs/random-samplings/2017/05/voting_in_america.html

[11] Ibid. Because 2016 poll worker figures were available in EAVS, we imputed values for 2014 by averaging figures for the two available years.

Methods & Results

First, we performed multiple ordinary least squares (OLS) regressions, using proportion of provisional ballots rejected as our target variable. We began with a one-feature model: median income regressed against the target variable. We proceeded to add more variables to each model, including the percentage of the county population with English fluency, the percentage of new registrations under the age of 25, and poll worker figures. The more variables we added, the higher the R-squared value, which is to be expected. We performed Likelihood Ratio tests at each stage to understand if our simpler models were the better choice. At each stage, the more complex model outperformed the restricted model. Our final R-squared value was .998, using eight features.

Concerned about multicollinearity, we employed, a variance inflation factor (VIF) test for each variable. Most variables were found to be collinear, particularly the percentage of the population identifying as white and median income. To address multicollinearity in our features, we pared down the final model to those features with low VIFs: total eligible voters, poll workers per voter, total provisional ballots submitted, and the proportion of new voter registrants for voters under the age of 25. However, the only significant variable was total eligible voters, which had a coefficient of zero. At this stage, we determined these features were not helping us further explain or understand provisional ballot rejection, exacerbated by our small sample size.

As a next step, we considered time series analysis, which would involve bootstrapping to generate provisional ballot rejection rates prior to 2010. However, as provisional ballots did not exist before 2004, it was deemed unethical to recreate data that has never existed. Consequently, we would not have enough data points for a time series.

Taking inspiration from a study of the effects on voting of Hurricane Sandy by Robert Stein[12], we developed a logit model comparing multiple state election administration and electorate features to high and low provisional ballot rejection. To do this, we created dummy variables based on if provisional ballot rejection was lowest in 2012, which became our new target variable, whether states issued an executive order allowing voters to vote in any precinct in 2012, and whether states allowed early voting and no excuse absentee voting.

To test the inference that removing the requirement to vote at designated precincts lowers the provisional ballot rejection rate, we constructed a logit binary classifier. We sought to isolate those features of election administration that contributed to a jurisdiction having the lowest provisional ballot rejection rate. We limited our feature set to variables available for all jurisdictions in EAVS.

Applying all features to the logistic regression resulted in high precision and recall of our model. Using Sci-Kit Learn's classification report to evaluate our model. We observed a stark difference in our model's ability to identify true positives of low ballot rejection rates. High ballot rejection rates had a recall of 0.95 where low ballot rejection rate had a recall 0.26. This is interesting due to a combination of a few features. New York City was the only region that had significant change in ballot rejection over time. It is also the only location that met these four conditions: the state was hit by hurricane Sandy, was declared a federal disaster zone (as opposed to Erie County, NY), offers limited early voting options, and lastly issued an executive order that removed a polling location requirement for voters. Due to these specific parameters, we can associate a low rejection ballot rate with the executive order. Of the nine jurisdictions with the lowest ballot rate in 2012, eight of which were in an executive order state, meeting the criteria for the association standard of causality.

[12] Stein, R. M.. (2015). Election Administration During Natural Disasters and Emergencies: Hurricane Sandy and the 2012 Election. *Election Law Journal, Volume 14, Number 1,* pages 66-73. DOI: 10.1089/elj.2014.0271

Conclusion

There is a clear correlation between low provisional ballot rejection rates and the removal of a polling location requirement. However, other aspects of election administration, most notably the number of poll workers per voter, prevents us from concluding that removing the polling location requirement rises to the standards of intervention and counterfactual.

Statement of Work

Christine Biddlecombe:

- Data Collection and Cleaning
- Linear Regressions
- Logistic Regression
- Paper Writing and Editing

Nathan Caplan:

- Data Collection and Cleaning
- VIF Calculations
- Logistic Regression
- Paper Writing and Editing

Ursula Kaczmarek

- FOIA Request
- Data Collection and Cleaning
- SHINY Application
- Paper Writing and Editing