How I learned to stop worrying and Love Voter Registration Files

A Thesis
Presented to
The Division of Mathematics and Natural Sciences and History and Social Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Theodore Dounias

December 2018

Approved for the Division
(Mathematics and Political Science)

___________________________ ___________________________

Andrew Bray Paul Gronke

Preface

This is an example of a thesis setup to use the reed thesis document class.

Table of Contents

List of Tables

List of Figures

Introduction

The democratic system is based on procedures as much as principles. The way that democracies chose to tally the will of the people is always a messy, controversial process. Thus the design and implementation of voting systems is far from being neutral; the decisions made on who votes, and how, when, and where they do so is inherently coupled with the outcome. Underlying those decisions is a nebulous, inconclusively answered question: are elections fair, and how can we make them more so.

The passage of the Help America Vote Act—or HAVA–(*HAVA*, 2002), which mandated states to update and consolidate public voter registration files, and created the US Elections Assistance Commission that makes available county level data, innovated the way we use data based approaches to answer this question. HAVA offered political scientists and statisticians direct access to the voting population's voting patterns, political registration, age, geolocation and much more; information that up to then was only accessible by sampling through surveys. The immense leap here happens because true population data does away with the need for sampling techniques that are often biased and inaccurate. We can now not only get a complete picture of the data, but also link and merge with other sources of information such as US Census data on religion, race, education, or income—work that has been lucrative for firms such as Catalist or Target Smart. By posing Political Scientific questions, and trying to respond with rigorous statistics, both disciplines tackle these data to face joint problems such as quantifying the quality of voter registration files (Ansolabehere & Hersh, 2010), or linking disparate voter records (Ansolabehere & Hersh, 2017).

Chapter 1

The State of the Literature

In this chapter I will go through the existing literature on Vote-By-Mail (VBM). I will define what Vote-By-Mail is; I will then summarize the expectations that researchers have of the effects of VBM on turnout, based on existing theories of electoral participation. I will continue with a summary of previous quantitative research on the effects that VBM and similar policies have had on turnout. I will conclude with some more general comments on the available data, and literature concerning the most commonly used quantitative methods.

1.1 What is VBM?

Gronke (2007, 2008), RMStein (1998)

Vote-By-Mail is a process by which voters receive a ballot delivered by mail to their homes. Voters then have a variety of options on how to return these ballots, ranging from dropping them off at pre-designated locations, to mailing them in, to bringing them to a polling place and voting conventionally. This varies across states that have implemented VBM. Some common forms of the VBM policy are:

- *Postal Voting*: All voters receive a ballot by mail, which can then be returned to a pre-designated location or mailed in to be counted. This is the current system in Oregon, is an option in Colorado, and is implemented by a number of counties in California, Utah, and Montana.

- *No-Excuse Absentee*: Voters can choose to register as absentee voters without giving any reason related to disability, health, distance to polling place etc. This is the case in 27 states and the District of Columbia.

- *Permanent No-Excuse Absentee*: This is similar to the previous system, but allows voters to register as absentees indefinitely, without having to renew their registration each year; they become de facto all-mail voters. This is in place in Washington, Kansas, and New Jersey.

- *Hybrid or Transitionary Systems*: In hybrid systems, voters receive a mail ballot but can choose to disregard it and vote conventionally. This is the case in Colorado. Transitional systems exist in states that have chosen to eventually conduct all elections by postal voting, but have given counties an adjustment period during which this shift is not mandatory, or mandatory only for certain elections. This is the case in California, Utah, and Montana.

Vote-By-Mail is also commonly considered a type of early voting, since voters receive their ballots around two weeks in advance of election day; they are also able to return that ballot whenever they wish within that time-frame. This means that Vote-By-Mail can be counted as a "convenience voting" reform. These are usually implemented by state and local governments with the argument that they either expand the democratic franchise by bringing in new voters, or by making it more likely that current registered voters participate in the electoral process.

1.2 The Calculus of Voting

Grimmer (2011), Burden (2013)

1.2.1 Why Turnout Matters

Geys (2006, 2016), Smets (2013) ++ book sources

Turnout is the most commonly used measure for electoral participation. It is important because it signifies the level of engagement of the population with the state, the level of incorporation of different subgroups of the population into democratic processes, and the legitimacy of elected officials. It is widely accepted that turnout should be maximized so that the democratic franchise represents the majority of citizens. Turnout for an election can be calculated or predicted, the difference being that in the former case we use data post-election to measure its absolute value, while in the latter we use a series of individual and community covariates to infer the levels of turnout for a future or past election.

Calculating turnout, at its core, involves the following equation:

$$\% \; Turnout = \frac{Total\ Ballots\ Cast}{Measure\ of\ Total\ Voting\ Population} \times 100\% \quad (1)$$

The choice of numerator is fairly obvious and universal; the denominator, however, is a different story. The two main statistics used are the total voting age population, and the raw number of registered voters in the geographical location we are examining. The total voting age population is used as a measure to incorporate the total amount of possible voters in a geographical area, and can be measured using data from the US Census. This causes some issues with voters that cross over to different districts; if someone lives in district A, it is still likely that they are registered to vote in district B. If this is not considered, the calculation of voting age population might be misrepresentative.

Using registered voters also brings with it two problems. First, the calculation necessarily occurs using voter registration files, which many times can include discrepancies, like deceased voters, voters included in multiple counties, or individual voters included multiple times. Furthermore, the total amount of actual voters among registered voters can be misrepresentative of democratic participation; consider that if a certain minority community has historically low registration rates, their lack of engagement will not be included in turnout rates, thus misrepresenting the level of inclusion in the district they reside in.

The punch line here is that how the turnout statistic is calculated is not a clear choice, and will have an impact on how studies are set up. To give one example, consider Oregon's Motor Voter program, that automatically registers voters when they interact with government services, like the DMV. It is conceivable that this reform will *decrease* turnout when measured as a percentage of the total registered voter count, but *increase* turnout when measured against total population. I will specify how I calculate turnout in the next chapter.**Need sourcing for this**

Statistical models of turnout can be constructed at either the individual or community level. At the individual level, a model is built to predict the probability of voting for every member of a group, and then sum over the members to create an estimate for turnout. Probit or Logit models are preferred. At the community level, researchers first choose a geographical level at which to calculate, which then constitutes the individual observation in the data that is used to create the model.

Both these models include a standard set of societal variables–at the individual and aggregate level–, policy variables–whether the district does Postal Voting, whether Voter ID requirements are particularly strict–, election-specific variables–closeness of election or campaign expenditure–and sometimes time-series data–previous levels of turnout–to make predictions on turnout levels. This type of analysis is not exclusively used to predict turnout but also to, as will be later shown, draw inferences on the effects that certain explanatory variables have on electoral participation.

Through meta-analyses on studies of turnout, it is possible to get a clear picture on what variables effect individual and collective choices to turn out. Three such studies are conducted by Geys (2006), Geys and Cancela (2016), and Smets (2013). Geys includes 83 studies of national US elections in his initial meta-analysis (Geys, 2006), later increasing that number to 185 (Geys and Cancela, 2016) and adding local elections. On aggregate-level models for national elections they conclude that competitiveness, campaign financing, and registration policy have the most pronounced effects, while on the sub-national level there are more pronounced effects for societal variables and characteristics of election administration (spending, voting policy, etc.). Smets and Van Ham (2013) examine individual-level predictors for turnout in a similar meta-analysis, and conclude that "age and age squared, education, residential mobility, region, media exposure, mobilization (partisan and nonpartisan), vote in previous election, party identification, political interest, and political knowledge" (Smets & Ham, 2013) are the most significant explanatory variables for turnout, along with income and race. ***I will add sources from books here***. I will specify the model I will use for turnout in the second chapter.

1.2.2 Theories of Voting

Aldrich (1993), Berinsky (2001, 2005), Edin (2007), Bendor (2003), Gerber and Green (2015), Matsusaka (1997), Fowler (2006)

Here I take one step back from turnout, and examine the theories surrounding individual choices to vote or abstain. There are three main theories outlined in the literature on why individuals chose to vote. While there is some overlap, the following are mostly distinct:

- *Decision "at the margins"*: In his 1993 study, Aldrich posits that voting is a low cost-low benefit behavior. Therefore, he continues, voting is a decision that individuals make "at the margins"; in most people, the urge to vote is not overwhelmingly strong, and therefore individuals will vote when it is convenient to them, when they are motivated by a competitive race, when policies are put in place to help them, and when they are subjected to GOTV (Get Out the Vote) efforts. For Aldrich, this is corroborated by the fact that most turnout models present consistent, yet weak, relational variables; if decisions are made "at the margins", then no single predictor would have an overwhelming result. This is also supported by Matsusaka (1997), and Burden & Neiheisel (2012). Matsusaka expresses support for a more "random" process of voting, where turnout models are ambiguous because of the difficulty that predicting "at the margins" entails (Matsusaka & Palda, 1999). Burden & Neiheisel (2013) also demonstrate support for Aldrich's thesis by using data from Wisconsin to calculate a net negative effect of 2% on turnout due to a similar slight shift in turnout.(Aldrich, 1993; Neiheisel & Burden, 2012)

- *Habitual Voting*: While Aldrich supports that there is no single overwhelming predictor of turnout, Fowler (2006) posits that future voting behavior can be strongly predicted using individual voting history. This leads to the conclusion that individuals are set to either be habitual voters, or habitual non-voters (Plutzer, 2002) by their upbringing and social circumstances, locking them into distinct groups. (Fowler, 2006)

- *Social/Structural Voting*: Close to habitual voting are those that support a model of social and structural voting; these researchers claim that the decision to vote or not is deeply rooted in socioeconomic factors, which means that the divide between traditionally voting and non-voting groups can only be bridged by directly dealing with the socioeconomic divide between them (Edin, 2007; (Berinsky, 2005). Their reasoning is that "at the margins" voting only addresses groups that do not face significant burdens against voting–like the working poor, or marginalized racial groups–, and are usually already registered. Similarly, they address habitual voting claims by arguing that they are too short-sighted; individuals themselves might be habitually voting, but their decision to do so is rooted in strong societal and policy factors.

Need more sources, will use books

1.2.3 How they Apply to VBM

Berisnky (2005), Banducci (2000), Gronke and Toffey (2008), and several applications of sources in above sections.

1.3 Previous Study Results

1.3.1 General Results

Arcenaux (2012), Bergman (2011), Burden (2014), Edelman + Pantheon Analytics (2018), Gerber (2013), Rhine (1995), Neihelsen (2012), Keele (2018), Richey (2008), RMStein (1997, 2007), Gronke (2007, 2008, 2012, 2017)

1.3.2 The Gerber Piece

Gerber(2017)

1.4 Voter Registration Files

1.4.1 Inaccuracy of Survey Data

Ansolabehere and Hersch (2012), Burden (2000), Deufel (2010)

1.4.2 The Importance of VRF

Books, mentioned later

I will include what I remember from some readings and what advisers mentioned in MATH241

As mentioned in my introduction, access to voter registration files has provided researchers with unique insight into the voting process. Quantitative research has expanded significantly, for three key reasons. First, VRF data exists in a consolidated, state-wide format at least for national elections. This means that the process of data collection involves interaction with significantly fewer government agencies, and a data wrangling process that can be quickly adapted to a set format. This is, of course, not to say that the process of data collection and handling doesn't still pose a significant challenge, as will become apparent in my second chapter. Second, there is a huge benefit attached to the fact that VRF data describes the whole population, rather than a sample. As mentioned in the previous section, survey data might give more insight into variables not included in VRF, but that comes at a steep cost for accuracy. Using VRF, the problem of self-reporting bias is eliminated for some studies, and transformed into a problem of record linkage and ecological inference for others (Ansolabehere & Hersh, 2017, Burden & Kimball (1998)). Third, wide public access means reproducibility and accessibility, which translates into greater accountability for researchers. This effect is important, even if mitigated somewhat by private data companies and access fees.

1.5 Common Methods Used and Problems Encountered

1.5.1 Methods

- *Synthetic Control Group*: Abadie (2010), McClleland (2017), Gronke (2017)
- *Record Linkage*: Ansolabehere and Hersch (2017), Harvey (1994, 97), Koudas (2013)
- *GLM (Probit/Logit/Poisson)*: Barreto (2004), Dow (2004)
- *DID*: Bertrand et al. (2002)
- *E.I.*: King (2013), Burden (1998), Calvo (2003), Chao (2004), Rm Stein (2002)
- *Mixed-Effects*: Gelman and Hill (2007)
- *General EDA and Models*: James et al. (2013), Chapman and Hall (2017)

1.5.2 Issues

Grimmer (2015) {Not always best to do inferential models}, Ansolabehere and Hersch (2010) {Problems with Voter Reg Files}, other sources from the previous section

Chapter 2

Hypothesis, Data, and Methods

2.1 The Data

2.1.1 Source

2.1.2 Structure

2.1.3 Wrangling

2.2 Hypotheses

2.2.1 Description of Hypotheses

2.2.2 Criteria

2.2.3 Expected Results

2.3 Methodology

2.3.1 EDA

2.3.2 Description and Parametrization of Models

2.4 Gerber Replication

Chapter 3

Results

3.1 EDA

3.2 Model Output

3.3 Gerber Expansion Results

Conclusion

References

Aldrich, J. H. (1993). Rational choice and turnout. *American Journal of Political Science*, *37*(1), 246–278. http://doi.org/10.2307/2111531

Ansolabehere, S., & Hersh, E. (2010). The quality of voter registration records: A state-by-state analysis. *Institute for Quantitative Social Science and Caltech/MIT Voting Technology Project Working Paper*. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/18550

Ansolabehere, S., & Hersh, E. D. (2017). ADGN: An algorithm for record linkage using address, date of birth, gender, and name. *Statistics and Public Policy*, *4*(1), 1–10. http://doi.org/10.1080/2330443X.2017.1389620

Berinsky, A. J. (2005). The perverse consequences of electoral reform in the united states. *American Politics Research*, *33*(4), 471–491. http://doi.org/10.1177/1532673X04269419

Burden, B. C., & Kimball, D. C. (1998). A new approach to the study of ticket splitting. *American Political Science Review*, *92*(3), 533–544. http://doi.org/10.2307/2585479

Fowler, J. H. (2006). Habitual voting and behavioral turnout. *Journal of Politics*, *68*(2), 335–344. http://doi.org/10.1111/j.1468-2508.2006.00410.x

Geys, B. (2006). Explaining voter turnout: A review of aggregate-level research. *Electoral Studies*, *25*(4), 637–663. http://doi.org/10.1016/j.electstud.2005.09.002

Matsusaka, J. G., & Palda, F. (1999). Voter turnout: How much can we explain? *Public Choice*, *98*(3), 431–446. http://doi.org/10.1023/A:1018328621580

Neiheisel, J. R., & Burden, B. C. (2012). The impact of election day registration on voter turnout and election outcomes. *American Politics Research*, *40*(4), 636–664. http://doi.org/10.1177/1532673X11432470

Smets, K., & Ham, C. van. (2013). The embarrassment of riches? A meta-analysis of individual-level research on voter turnout. *Electoral Studies*, *32*(2), 344–359. http://doi.org/10.1016/j.electstud.2012.12.006

The help america vote act of 2002, Pub. L. No. HR3529 (2002).