Turnout and Mail Voting in Colorado; or How I Learned to Stop Worrying and Love Voter Registration Files

---

A Thesis
Presented to
the Interdivisional Committee for Mathematics and Natural Sciences,
History and Social Sciences
(Mathematics and Political Science)
Reed College

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

Theodore Dounias

December 2018

Approved for the Committee
(Mathematics and Political Science)

Paul Gronke	Andrew Bray

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–(Robert Nay, 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. Public voter files were initially used by private corporations like TargetSmart or Catalyst, which cleaned up the files for use by political campaigns trying to tailor their message as closely as possible to individual voters. Researchers quickly realized the massive potential that such data has, and started partnering with such firms or conduction independent cleaning and structuring of such data themselves. These data provided information at the individual level, with geocoding usually at the precinct level. This in turn meant that, though very complex as a task, it became possible to use such files along with census data at the block level to estimate individual characteristics like race, education, or income at incredible levels of accuracy. Even without the resources of a multi-million dollar private market, such methods along with voter files alowed researchers to make concrete inferences of individual characteristics (E. D. Hersh, 2015). Voting related theories derived from political science are now commonly tested using advanced statistical methods and huge amounts of data; 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 first go through some general literature on theories of voting decisions. 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.

1.1 Deciding to Vote

1.1.1 Why Turnout Matters

Turnout is the most commonly used measure for 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. Turnout is a metric for how widespread democratic participation is; it is one of the best quantitative measures of the strength of the democratic franchise, alongside qualitative metrics such as voter education and information. Turnout for an election can be calculated or predicted, the difference being that in the former case we use data post-election that is reflective of the actual number of voters, while in the latter we use a series of individual and community covariates to infer the levels of turnout.

To calculate turnout, we simply divide the number of ballots cast by the potential voting population, as in 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 three main statistics used are the total voting age population, voting eligible population, and the number of registered voters in a certain geographical location. The total voting age population–all individuals over 18 years of age–can be measured using data from the US Census. However, such an interpretation of voting age population positively counts individuals of age that are not allowed to vote–people

with severe mental illnesses or felons–, and does not count oversees voters or military personnel. Michael McDonald offers an alternative to voting age population he calls "voting eligible population", which corrects for such individuals (McDonald, n.d.).

Counts of registered voters are also a useful tool for calculation of turnout, as they usually require no estimation. These counts can simply be extracted from voter registration files. Using registered voters, however, also brings with it two problems. First, voter registration files 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 they are not included at all in calculations of the turnout statistic.

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. This happens if more people register to vote, but do not actually do so–in other words, both number of registrants and number of ballots cast are increasing, but the former increases at a larger rate than the latter. I will specify how I calculate turnout in the next chapter.

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 specify the model I will use for turnout in the second chapter.

1.1.2 Theories of Voting

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 following the expansion of early voting access in the state. (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 (Berinsky, 2005; Edlin, Gelman, & Kaplan, 2007). 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.

- *Resources and Organization*: To some extent growing from structural theories of voting, resources and organizations theory emphasizes the interaction of personal political and societal characteristics of voters, and actions taken by politicians to mobilize participation. This theory is very broad in the inputs it assesses for voter participation, ranging from practical issues of access and resources–how easy it is for someone to vote, if they so choose–, to public policy feedback effects and signaling–how the government's policies effect the people, and how they react to them–, to how political parties and groups choose to mobilize and approach voters (Rosenstone, 2003). Apart from Rosenstone and Hansen's work (2013), there have been several studies examining voter participation based on resources and organizations theory, a lot of which come from the public policy side of political science. Some examples are Chen's study of how distributive benefits like federal emergency aid affect participation among recipients, after controlling for partisan characteristics (2012), or Mettler and Stonecash's examination of correlation between welfare program participation and political mobilization (2008), or Campbell's analysis of social security recipients and their voting patterns (2002). The punchline in all these studies is that public policy is correlated with trends in participation, either because recipients of benefits wish to protect such programs, or because of the interaction between partisanship and government support, or because of access related to resources and voting laws (Campbell, 2002; Chen, 2013; Mettler & Stonecash, 2008).

1.2 From Theory to Policy

1.2.1 Voting Methods

I have already flagged in my introduction the reason why theories behind voting choice matter: each construct an image of the electorate that reacts differently to policy change around voting. They are all an answer to the fact that elections policy, and how we conduct elections, is not value neutral but has implications for turnout, which in turn has implications on the franchise of democracy.

In trying to respond to the issues set up by theoretical paradigms, different states–both in the world and US contexts–have adapted to different ways of conducting elections. In the US, voting styles can be simplified into three categories:

- *In-Person Election Day*, for which all individuals are required to vote at a polling place, on a single election day. There can be some leeway for overseas voters, or excused absentee voters, but the vast majority of people will have to be present to vote in a particular time frame.

- *In-Person Early Voting*, for which all individuals must vote in person at a polling place or vote center, but the timeframe for voting extends for around two weeks, not a single day.

- *Vote-By-Mail, Absentee Early Voting*, for which individuals have a clear, no-excuse-necessary option for not being present when they vote, or for filing in a mailed ballot and droping it off at designated locations.

For the purposes of this thesis I will examine the latter category, and specifically Vote-By-Mail. The reason behind this is that the model of in-person, election day voting is usually seen as the baseline, the "vanilla" way of conducting elections if you will. Therefore it has been of interest for researchers to examine if other systems can outperform that baseline. Specifically, it is most interesting to examine voting styles that are heralded for their expansion of turnout, to see whether popular beliefs on their benefits and drawbacks hold; if they are different from the base model of conducting American elections, or if they present new challenges and unique selling points. Vote-By-Mail is particularly interesting because it is quickly taking the form of a trend in state elections, as more and more states are enforcing more open models of VBM. In the next section, I will more closely examine the particulars of Vote-By-Mail.

1.2.2 What is VBM?

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–the two first options are most commonly implemented, with a very small number of states still operating polling places for mail ballots. 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. All-mail elections currently occur in Oregon, Washington, and Colorado.

- *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 a very large number of the no-excuse absentee voting states like Utah, California, Montana, Arizona, New Jersey, and others.

- *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 (State Legislatures, n.d.).

1.2.3 How Theories Apply to VBM

Under Aldrich's paradigm, vote by mail would not effect significant change in voting behavior. The whole concept of a decision "at the margins" is that the forces at play when an individual decides to vote are overwhelmingly strong both ways, so any effect that policy can have will minimaly shift these margins. If, for example, we take a presidential election the forces at play include the media, national committees, social effects etc. In this environment some added convenience does not significantly add to an individual's decision to turn out. However, this would indicate that at a local level, where national and media effects are less strong, the effect of VBM on turnout might be more significant. The effect would be present for all groups, not only those currently registered, since voting would be easier uniformly.

If we asume habitual voting, the conclusion on VBM would differ significantly. In this case, the effect to be considered is how VBM impacts already formed habits around voting. It could be argued that VBM has no effect, which follows if we assume that voting habits formed do not shift if the mode of voting changes. It could also be argued that VBM might have a negative effect on turnout in the short term, because it disrupts the habit of election day for a readjustment period, before people settle into new groups of habitual voters and non-voters, adapted to the new policy context.

Under social and structural voting contexts, VBM retains rather than stimulates new voters (Berinsky, 2005). This means that already registered and semi-active voters are more likely to participate, but there is no significant change in the amount of new voters entering the franchise. This would mean that traditional forms of voting policy that emphasize access to the polls will do nothing to bring in disenfranchised people, and potentially hide the problem under an inflated turnout statistic calculated on registered voters. Berinsky in particular emphasizes the need for a shift towards voter education, rather than early voting or VBM policies (Berinsky, 2005).

Vote-by-Mail is obviously not a welfare or spending program, but it does expand individual resources in terms of voting capacity. A ballot delivered to your home means that less resources need to be expended in the act of voting, which in turn has both a practical effect–building capacity–and a more behavioral effect–a feeling of inclusion, an interaction with the process of voting that comes through a ballot at your doorstep that would not exist if you had not gone to a polling place (Schneider & Ingram, 1990). Under a resources and organizations framework, both these effects are most likely to be net positive to political participation, and as such would predict a strong, positive effect of VBM on turnout.

1.2.4 General Results on VBM

I will start with studies that show a negative effect on turnout. Bergman (2011) uses a series of logit models of individual voting probability in California, during a period where part of the state conducted VBM elections, while others maintained traditional voting. This is called a "quasi-experiment", and is frequent throughout the literature. Bergman's results show a statistically significant drop in voting probability in VBM counties (Bergman & Yates, 2011). Using a similar method, Keele (2018) takes a single city in Colorado, Basalt City, which is divided into two different voting districts using different voting systems. The conclusion is, again, a 2-4% drop in turnout along the VBM part of the city (Keele & Titiunik, 2017). Burden et al. (2014) takes a different approach, using country-wide election data from 2004 and 2008 presidential elections, and compares districts based on early voting practices. Their results show a significant drop in turnout, which can be associated to VBM as well due to its closeness to EV (Burden, Canon, Mayer, & Moynihan, 2014).

In contrast, Gerber et al. (2013), applying both individual and county-level models for the state of Washington, reach the conclusion that VBM increases turnout by around 2-4%; they use the same quasi-experimental model that offers itself to researchers in states that are under transitional systems (Gerber, Huber, & Hill, 2013). R.M. Stein also reaches a similar conclusion when examining Colorado's practice of "vote centers", which are non-precinct attached polling places, which can service multiple counties (Stein & Vonnahme, 2008). I include this paper here due to the link that voting centers have with VBM, as they serve as drop-off points for mail-in ballots. Richey (2008) examines the effects that Oregon's VBM program has on turnout by using past elections data, concluding a 10% positive trend associated with the policy (Richey Sean, 2008). This effect is studied again by Gronke et al.(2012) who find a similar positive effect with much lower magnitude, which might point to a novelty effect: the existence of diminishing returns in turnout after the implementation of this policy (Gronke & Miller, 2012). Gronke et al. (2017), again studying Oregon but focusing on Oregon's Motor Voter program, find evidence of positive association to turnout [@]. I include these effects due to Oregon being an exclusively VBM state, and because this paper uses a "synthetic control group" model, a particularly interesting statistical technique. Lastly, I include a study conducted by Pantheon Analytics on Colorado, which compares actual turnout to predicted levels for VBM counties in Colorado. The results show a positive effect of approximately 3.3% due to VBM (Edelman & Glastris, 2018).

The conclusion to be drawn from this section is that results on VBM vary significantly. There are multiple studies, using multiple methods, on multiple states, with multiple results. This only adds to the importance of being careful when constructing models and hypotheses to test VBM's effects on turnout, as assumptions made in the process can critically impact the results.

1.3 Voter Registration Files as Data Sources

Before concluding this chapter, I want to briefly discuss some background research into the use of voter registration files for the purpose of elections research. This may seem like an abrupt shift from the previous section but, as I mentioned in my introduction, access to such files is what motivated and facilitated a lot of the aforementioned studies in the first place. I will not directly go into the structure of such files; such a section will be included later on in this thesis.

1.3.1 Inaccuracy of Survey Data

Apart from Voter Registration Files, the main source of data on the American electorate is national surveys, like the American National Election Studies' survey (ANES), or the Cooperative Congressional Elections Study (CCES). These are post-election surveys, distributed to voters, which include fields associated directly with voting–participation, precinct, which party you voted for–and indirectly, through questions on individual characteristics like race, income, or gender. On the surface these seem like a better source of data, since no record linkage or ecological inference need be made to connect individual voters with an extensive list of covariates. There is, however, a significant problem with these data: survey misreporting (Burden, 2000).

A cursory glance at the CCES and NES estimates of turnout reveals the existence of a problem right off the bat: turnout calculated through surveys is usually higher than reported figures. When looking at surveys a bit closer, using either private, extensive data files like Catalyst (Ansolabehere & Hersh, 2012) or validated voter files from the late 20th century (Deufel & Kedar, 2010), the results show consistent misreporting among certain groups, that tend to either be politically engaged non-voters or minorities and low socioeconomic status individuals. This gap, according to Deufel et al. (2010), has served to propagate societal stereotypes and class entrenchment into studies on turnout, which in turn negatively effect policy, since research using the ANES and CCES are widely used to study turnout among the groups that are consistently misreporting. Admittedly, the fact that misreporting happens among specific groups does open the way for statistical methods to compensate for the bias introduced, but for the purpose of my thesis I will prefer the use of Voter Registration Files.

1.3.2 The Importance of VRF

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 reproducible and accessibility, which translates into greater accountability for researchers. This effect is important, even if mitigated somewhat by private data companies and access fees.

Chapter 2

Hypotheses and Methods

In this chapter, I introduce a series of questions resulting from the literature review of Chapter 1, which I will use to formulate hypotheses. I will then operationalize these hypotheses, and attempt to predict analytical outcomes based on the theories of Chapter 1. Following these hypotheses, I will outline key methods I will use to test them.

2.1 Hypotheses

2.1.1 Questions

Before moving in to outlining hypotheses, the first step necessary is to frame a series of questions, which the hypotheses will flow from. Based on relevant research, the most obvious first question to ask would be:

Q1: *What is the effect of mail voting on turnout?*

I went through this question substantially in the previous chapter; it should be clear that depending on which paradigm of participation choice is present, the answer here can be radically different. In order to best answer the previous question, it is necessary to establish some conditions on importance of effect. Therefore it is also necessary to ask the following question:

Q2: *Does this effect persist when accounting for other relevant predictors of turnout?*

The last question asked in this thesis is more specific to a particular formulation of Aldritch's hypothesis on voting "at the margins". I mentioned in the previous section that VBM could be theorized to have a more significant effect when discussing elections at the local level, or the regional level, rather than national general elections. Therefore a third question is:

Q3: *Is the effect of VBM on turnout more pronounced as significant, national determinants become less strong?*

2.1.2 Hypotheses

Using the above questions I can now move on to formulate more clear hypotheses. Before diving right into that, I note that I intend this thesis to serve two purposes: first, to test voter choice theories between each other; second, to serve as an analytical tool for later evaluations of mail voting as policy. Based on the theoretical review of the previous chapter it should be apparent that of these two purposes, the former is primarily addressed, with the later tangentially arising from my conclusions. The hypotheses in this section spring mostly from a wish to test theories of voter choice, and in particular a wish to defend the theory of voting "at the margins" as introduced by Aldritch. Therefore all hypotheses in this section will be phrased from the perspective of this theory, with the competing alternate hypotheses being counter-claims potentially rooted in different theories of voter participation.

In response to Q1, Q2, a first hypothesis is:

H1: *Mail voting is another marginal effect on voting decisions, and therefore does not significantly affect turnout*

The alternative hypothesis would be:

H1': *Mail voting significantly affects turnout, even compared to other metrics*

Similarly, for the third question, a corresponding hypothesis derived from Aldritch's paradigm is:

H2: *The effect of VBM on turnout is more pronounced as national effects dull*

The alternative hypothesis is:

H2': *The effect of VBM on turnout is consistent and independent*

2.1.3 Criteria

A first, glaring issue that needs to be clarified is the apparent contradictions between my two hypothesized results. This becomes clear, however, if I define "significant effect" in the context of my first hypothesis. Aldritch's paradigm does state that "conveniences" like mail voting should not have significant effects, but those effects are defined in the context of huge, clashing forces that vastly outweigh them. This does not necessarily mean that they are literally non-existent, but that they are poor indicators of consistently increased turnout. Therefore, I will confirm my first hypothesis not only if the effect of mail voting on turnout is statistically insignificant, but also if it is relatively small in comparison to the effects of other variables I include. I will confirm the alternative hypothesis if, across multiple of the models I will parametrize and fit, VBM retains a consistent, significant effect on turnout. If the effect is negative, this may point to a habitual or structural voting paradigm being present. If the effect is positive, this may be a signifier that issues of convenience in voting–having a mail

delivered ballot, voting from your kitchen table etc.–have a particularly strong effect in the examined elections.

Moving on to the second hypothesis. It is extremely hard to correctly operationalize and account for all variables going into turnout. Therefore, instead of trying to include all national effects into a model and try to see how they interact with VBM, I will test my hypothesis on more localized elections. At least in theory, I can assume that if mail voting significantly impacts people's decision to vote, it will be in a context where the convenience of voting significantly outweighs information effects from national media, communal pressures, or national campaigns. This can be found to some extent in primary elections, but much more significantly in off year local state elections. A potential re-formulation of the second hypothesis, that makes it more specific to the criteria I have set, is:

H3: *The effect of VBM on turnout is more pronounced in local or off year elections*

I will confirm this hypothesis if mail voting has significantly larger positive effects on turnout in smaller, local elections.

2.1.4 Importance of Hypotheses

The importance of these hypotheses is intrinsically tied to the importance of different theories of electoral participation. Confirming or rejecting each hypothesis–even when only applied to a single state–serves as an argument for or against one of the aforementioned theories. The theories in and of themselves are significant, since they form a part of a broader literature on elections, democracy, and electoral processes, that can be said to be foundational to political science as a whole. Elections are the root from which all democratic governing springs; understanding why people participate in them is understanding how they choose to be included or excluded from the process of policy-building, and how they interact with the state.

Additionally, from a public policy perspective, these hypotheses are significant since they serve as metrics for the effectiveness of mail voting as an electoral reform. Whether, in general, mail voting increases turnout is directly connected to whether it is successful in expanding the democratic franchise. If it is not, questions can be raised as to the effectiveness of expanding voter access through elections administration, rather than education, or even measures like voting-day-holidays or local transportation to poling places. In local elections in particular, significant effects of mail voting could be precursors to more general involvement of individuals in their local politics. This may open the way to numerous comparative studies on local politics between states that apply VBM and states that do not.

Lastly, from a narrower perspective specific to the study of early and mail voting, my first hypothesis can still be said to be significant, yet mundane. It does its job according to the particular state I chose to look at–in this case Colorado–to add to existing literature on mail voting effects in different parts of the country. However, my second and third hypotheses are much more unique in their scope. There have not been many studies that look at VBM at a more localized level, and any addition to the literature on this front–however limited–could be significant.

2.2 Methodology

Before directly defining all parameters of the models I will later use in writing this thesis, I will go through each type of method to provide some background on the statistics behind the models. In the next chapter, I will introduce the data and fully outline my models. This section should serve as a general introduction to the methods. I will not extensively go through the statistics behind linear or multiple regression, but will assume that it is common knowledge. For an extensive introduction to such methods, James et al.(2017) or Chihara and Hesterberg (2011) are particularly useful.

2.2.1 Logistic Regression

Let function $f : [0, 1] \to \mathbb{R}$ be defined as:

$$f(p) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

This is called the logit function or, when p refers to a probability, the log-odds function. When modelling a binary response Y, which follows a Bernoulli distribution:

$$Y \sim \text{Bernoulli}(p),$$

the logit function can be used as a link function to model Y in a generalized linear model. The generic form of a generalized linear model looks like:

$$f(Y) = XB,$$

where Y is a vector of response variable values, X is a matrix of predictors, and B is a matrix of coefficients to be estimated. The function f is called a link function, because it "links" the response variable with the set of predictors included in the model. This is typically done to ensure that the range of values outputted by the model are consistent with the range of the response variable. When wanting to compute a model on a binary response through its corresponding Bernoulli distribution probability parameter, the inverse logit function should be a perfect fit for a link function, since it maps values from all real numbers to a range between 0 and 1. Using the inverse logit function, we arrive at the final form of logistic regression, which is:

$$\mathbb{P}(Y_i = 1) = \text{logit}^{-1}(XB)$$

Conveniently, despite the use of a link function, there is an easy way to interpret the coefficients of such a regression. While obviously individual values from the B matrix will not be particularly helpful, e^B can be used as a matrix of multiplicative, one-unit shifts in the value of the probability that $Y_i = 1$. This means that a one unit increase in any predictor will cause an effect equal to multiplying p by the exponent of the corresponding coefficient. (James, Witten, Hastie, & Tibshirani, 2017)

2.2.2 Generalized Additive Models

In simple logistic or linear regression, there is an assumption made on the functional form of the relationship between predictors and response variable. These are called parametric models, where the data is exclusively used to estimate values for coefficients. Non-parametric models, on the other hand, use the data to estimate both coefficients and the function that serves to connect response to predictors. While on the surface this seems like a great idea (more reliance on your data and less assumptions!), such an exclusively non-parametric model would suffer greatly from the curse of dimensionality– where the addition of multiple predictors or over-reliance on data leads to substantial over-fitting.

The solution, then, is a Generalized Additive Model, or GAM. This model lets us fit a different functional form to each observation, allowing for assumptions to be made on the data where it is safe to do so, and for non-parametric fitting when it is necessary. This model looks like:

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j f_j(x_{ij}),$$

where y_i the i-th response variable, α is the intercept term, f_j, β_j a series of p functions and coefficients, and x_{ij} the i-th observation for the j-th predictor. Note that for $f_j(x_j) = x_j$, this is a multilinear regression! (James et al., 2017)

A type of most commonly fit functions–and the type I will make use of–are smoothing splines. These are cubic functions connected at specific points called "knots", with the limitation that the full function must be continuous and smooth. These are particularly useful when modeling time variables, as they can be fitted to variables like years or months in order to distinguish a secular trend from a general trend over time. In terms of this thesis, this will help when responding to Q2 as it was framed earlier in this chapter (Barr, Diez, Wang, Dominici, & Samet, 2012).

2.2.3 Multilevel Models

Multilevel models–otherwise known as hierarchical or "mixed effects" models–can be intuitively pictured in two ways: either as a set of models working on different "levels", where one is calculated first, with its effects having implications for the second, or as a model where some of the parameters estimated act under a particular series of constraints. Multilevel models are, in essence, a compromise between levels of "pooling" data. If the dataset on which parameters are being estimate operates in different units of observation–say on the individual and county level–you could run a model that treats all individuals as coming from the same larger group; this would be a complete pooling model. You could also add indicator variables for each and every group, de facto estimating n different models for n groups; this would be a no pooling model. Multilevel modelling offers partial pooling (Gelman & Hill, 2006).

To consider what this model looks like, let's assume a dataset comprising of a vector of values for the response variable Y, a matrix of i individual level predictors X, a matrix of j group level predictors U, intercept terms α, individual level coefficients

B, and group level coefficients Γ. Based on this, a multilevel model with intercept terms varying by group looks like:

$$Y_i = \alpha_{[i],j} + X_i B \ , \qquad \alpha_{[i],j} \sim N(U_{j[i]}\Gamma, \sigma_\alpha^2)$$

Multilevel models can be fit using the `lme4` *R* package that uses restricted maximum likelihood calculations for estimating coefficients (Bates, Mächler, Bolker, & Walker, 2014). Multilevel modelling also works perfectly well with general additive models! In *R* this can be accomplished with the `gamm4` package (S. Wood & Scheipl, 2017).

Chapter 3

Case Selection, Data, Model Parametrization

In this chapter, I will first go through a description of the state of Colorado; its demographics, its politics, and its selection for the purposes of this thesis. I will then go through the sources and wrangling of the data I obtained on Colorado's elections. Finally, I will fully define the models I will be using to test the hypotheses outlined in the previous chapter.

3.1 The Centennial State and Its Voters

3.1.1 Demographics

Colorado–named the Centennial State due to assuming statehood on the centennial of the Union–lies in the Southwestern United States, with its Western half squarely atop the Rocky Mountains. Based on its estimated population of just over 5.5 million, Colorado is the 21st most populous state, and ranks 37th in population density. The vast majority of that population is gathered in a series of urban areas that comprise a North-to-South strip in the middle of the state, containing the Denver-Aurora-Lakewood Metro Area, Colorado Springs, Pueblo, and Fort Collins. Apart from the Western town of Grand Junction, the rest of the population resides in vast rural areas.

Colorado is landlocked, and far from any coastal town; in place of seaside resorts, Colorado attracts a substantial amount of tourists to its mountains every year. Therefore the more mountainous regions have developed skiing and mountaineering resorts. They also heavily depend on federal money and protection for national parks. These importance of these characteristics will become apparent in the following sections.

Colorado has a median age of 34.3 and median household income of $65,685. Colorado's population is mostly white, with a higher minority group population density in its Southern regions, as shown in figure 3.1. (Bureau, 2010) The conclusion here is that Colorado is a relatively young, mostly white, and fairly well-off state that is increasingly getting more diverse, particularly in the South. These factors are important as they serve to associate Colorado with other states; such associations are

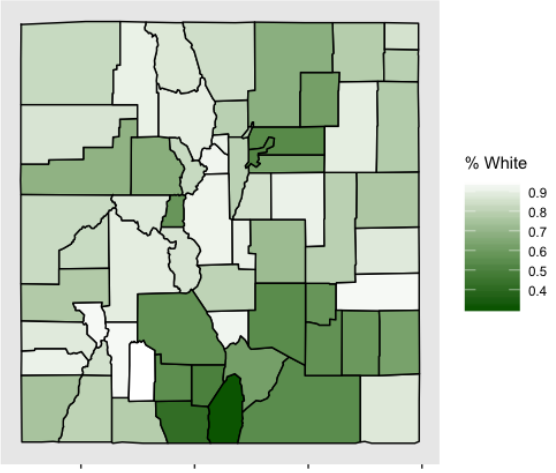

Figure 3.1: White voters per Colorado county

useful for the replication of this study or the generalization of my results.

The State Capital is Denver; Colorado is split into 64 Counties, of which the most populous are, in no particular order, the following: El Paso, Denver, Arapahoe, Jefferson, Adams, Larimer, Boulder, and Douglas. These counties comprise 73% of the total population of Colorado.

Table 3.1: Colorado population for largest counties

County	Total Population	CO Population %	Largest Metro Area
Adams	441603	0.08781	Denver-Aurora-Lakewood Metro Area
Arapahoe	572003	0.1137	Denver-Aurora-Lakewood Metro Area
Boulder	294567	0.05857	Boulder
Denver	600158	0.1193	Denver
Douglas	285465	0.05676	Denver-Aurora-Lakewood Metro Area
El Paso	622263	0.1237	Colorado Springs
Jefferson	534543	0.1063	Denver-Aurora-Lakewood Metro Area
Larimer	299630	0.05958	Fort Collins
Other	1378964	0.2742	
Colorado	5029196	100	

3.1.2 The Politics of Colorado

Curtis Martin (1962) notes that Colorado, due to its status as a frontier state, has always been fiercely democratic and independent. He connects this fact with Colorado's past, by pointing out that its political institutions were deeply rooted in mining culture, ordinary citizens' participation,a strong feeling of being "far away" from sources of centralized power in the coasts, and a wish for the protection and preservation of Colorado's natural environment. As such, Colorado can be described as a populist state with a strong libertarian streak, that highly values democratic processes when

This 1964 study of Colorado politics rings true to this day. One needs not search for long to see instances when Colorado honored this description. One example is TABOR, or the Taxpayer's Bill of Rights; a strongly libertarian, small-government, populist series of regulations that mandated a referendum for any measure that would increase state taxes, and caped government spending. TABOR was passed by referendum in 1992, and later amended in 2005 after the dot com economic crisis exposed the fact that inability to spend is very bad for a state government trying to jump start its economy. (Assembly, n.d.)

Similarly, Amendment 64 passed in 2012 made Colorado one of the first states to legalize the selling, possession, and consumption of recreational marijuana–a policy advocated by progressives and libertarians alike. Colorado was also the staging ground for what has been coined the "Sagebrush Rebellion": a movement primarily consisting on ranchers in dispute with the federal government over land use laws and wildlife protection. While this "rebellion" primarily consisted of battles in local legislatures or elections in the 1970s, its echoes can be heard till today in events like the Bundy Standoff, with ranchers taking up arms against federal employees and occupying federal land (Thompson, 2016).

Setting policy aside, this description of Colorado is also confirmed by polling data and election results. While being traditionally more conservative, inflows of immigration from the South coupled with increasing urban liberalization and tourism has led the state from leaning republican to being aggressively purple: the quintessential swing state. Colorado voted both for and later against Bill Clinton, voted for G.W. Bush twice, and has supported democratic presidential candidates since (Hamm, 2017). The trend is also, maybe surprisingly, consistent when considering both rural and urban voters; the divide that is said to plague other states seems to have passed by Colorado. Additionally, when polled on trust of federal or local governments, Colorado residents are systematically skeptical; in a random sample poll conducted by Cronin and Loevy (2012) in 2010, 56% stated that their state officials were lazy, wasteful, and inefficient. However–again indicating a libertarian, independent streak–most Coloradoans from 1988 to today consistently believe that their state is "on the right track."[1]

3.1.3 Voting in Colorado

Each County individually administers local, coordinated, primary, and general elections, under the supervision of the Colorado Secretary of State. This means that each county individually handles the voters registered in that county. Unsurprisingly, the same eight most populous counties are also the counties with the majority of registered voters, as their registrants comprise 73% of total Colorado registered voters (as of November 2017). As table 3.2 shows, these eight counties have a registration rate between 60-80%, compared to a Colorado-wide rate of about 67%. Registration rates for all counties are also graphically depicted in figure 3.2. In terms of Party registration, Colorado as a whole leans democratic by a very narrow margin (figure 3.3).

[1] Colorado College Citizens Polls, taken from Cronin et al. (Cronin & Loevy, 2012)

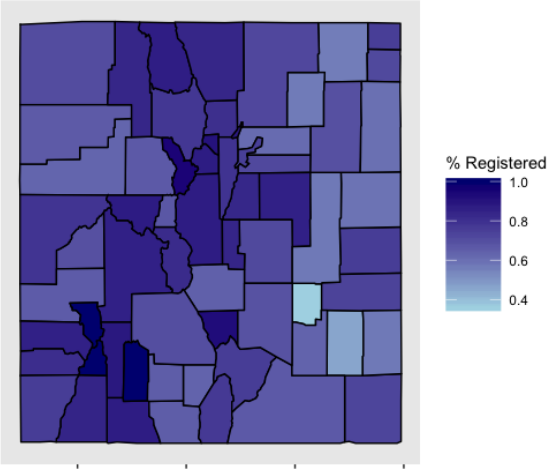

Figure 3.2: Registration rates per Colorado county

Table 3.2: Colorado voter registration for largest counties

County	Total Registered Voters	County Voter Registration Rate	% of Statewide Registrants
Adams	270,303	0.61	0.07
Arapahoe	410,546	0.72	0.11
Boulder	237,091	0.80	0.06
Denver	450,616	0.75	0.12
Douglas	237,659	0.83	0.06
El Paso	445,708	0.71	0.12
Jefferson	422,362	0.79	0.11
Larimer	250,626	0.84	0.06
Other	1,009,392	—	0.27
Colorado	3,734,303	0.67	1.00

In the past 25 years, there have been a series of key changes in the way Colorado administers elections, in relation to Vote By Mail and other reforms targeted and expanding the democratic franchise. In 1992, Colorado introduced no-excuse absentee voting, allowing voters to either physically pick up a mail ballot at a Vote Center or County Office, or have a ballot mailed to them prior to election day. In 2008, this reform was expanded to a permanent Vote-By-Mail system, which gave voters the option to be permanently put on a list of addresses that received mail ballots prior to the election. The State also entered a transitional status to full mail elections, giving counties the option to make all coordinated local elections, general elections, and primary elections exclusively VBM. In 2013, the Colorado State Legislature passed HB13-1303: The Voter Access and Modernized Elections Act, which mandated that every voter currently registered receive a mail ballot for all future elections. The Act also expanded the use of Vote Centers instead of traditional polling places, instituted same-day voter registration, and revamped the way active and inactive voter status

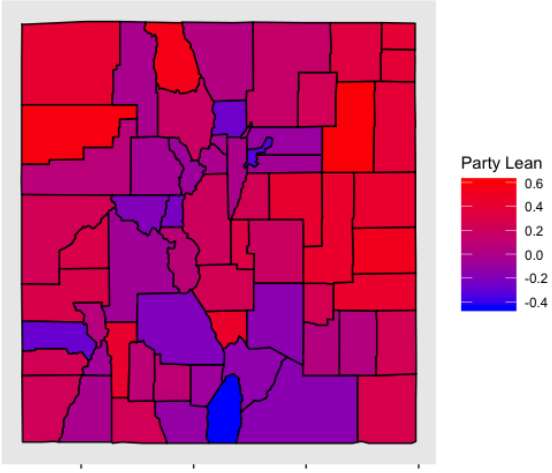

Figure 3.3: Democratic/Republican party lean per Colorado county

was designated on voter rolls–more on this in future sections. These changes are summarized in Table 3.3.

Table 3.3: Key changes to Colorado elections policy

Year	Key Changes
1992	No Excuse Absentee Statewide Implementation
2008	Permanent No-Excuse VBM Lists, Option of Full-VBM Elections
2013	Automatic Mail Ballot System Implemented Statewide, Established Vote Centers

3.1.4 Colorado as a Case for this Thesis

Colorado presents such an interesting case for research on Vote By Mail exactly because it has gone through such a long transitional process to reach its current elections system. It has steadily developed voting policy through a mixture of state mandates, county action, and outside policy motivations. Colorado's streak of independence and direct democracy is also very apparent in this shift in electoral practices, since they have been passing policies trying to expand participation for a very long time. It gives researchers access to approximately 22 years during which at least part of the state conducted elections partially by mail, making comparative, county- or individual-level case studies particularly alluring. Colorado's streak of independence and direct democracy is also very apparent in this shift in electoral practices, since they have been passing policies trying to expand participation for a very long time.

On a more general level, Colorado is interesting exactly because it is "typical" but with a wild streak. It is typical rocky mountain country, great planes country, and liberal urban city but all *in one state*. In is libertarian yet increasingly Democratic. It heavily relies on state funding for national parks, yet rebels against federal land use laws. It is a frontier state with traditional values, that overwhelmingly supports marijuana legalization. It is also a consistent purple state, with a Democratic Governor and House, but Republican Attorney General, Secretary of State and senate. This means that Colorado is a combination of distinct national effects, but also local effects that make it significantly different from national trends as a whole. In this environment, predicting results of policy can be difficult, but extremely salient as

- **County and individual level voting data**: turnout, party registration, total registrants
- **Information on individual elections**: date, ballots cast, voting methods, county, election descriptions

In the process of my research, I have acquired sufficient data to cover the second and third of these areas. I was unable to procure individual level data on demographic characteristics apart from gender, age, and party registration. However, reasonable conclusions can still be drawn from county or precinct aggregates.

3.2.1 Sources and first glance

I used two sources of data: Colorado voter records procured from the Colorado Secretary of State's office, and demographic data from the 2010 US Census. In the process of procuring these data I was aided by a series of other researchers and professionals with experience in the field of elections administration; they are mentioned in my acknowledgements.

2010 US Census

The US Census is conducted country-wide every ten years, with the goal of procuring accurate data on the demographic characteristics of the population. The Census uses a combination of federal field workers conducting door-to-door canvassing and statistical methods for data aggregation. From the 2010 Census–which is publicly available online–I get total population counts, characteristics on race, and rural/urban population counts for Colorado.

I use two datasets from the Census. For both, the unit of observation is one of the 64 counties of Colorado, and both include the same total population counts. One contains racial demographic characteristics and the other contain percentages of rural and urban populations in each county. The racial demographic dataset needed some wrangling work to extract a percentage of white residents in each county. Individuals were coded as "white" when they identified as exclusively white–this doesn't include mixed-race individuals reporting white ancestry.

Colorado Voter Files

As any state, Colorado maintains a statewide registry of all currently registered voters. This registry is typically under the purview of the Secretary of State–in this case, Wayne W. Williams. Voter Registration Files are constantly updated with new information on existing voters, new voters, or with the removal of inactive or otherwise ineligible voters. Therefore, this file will be different every time it is accessed or shared. Based on when this file is accessed, only a "snapshot" of the file can be obtained. I have managed to procure "snapshots" for each year between 2012 and 2017.

Similarly with VRFs, a Voter History File is maintained and constantly updated by the state. This file is uniquely connected to its VRF: only voters showing up as

registrants will have their histories included. I have similarly procured "snapshots" of the Voter History File for the years between 2012 and 2017.

In the Voter Registration files, the unit of observation is the individual voter, and all variables are initially coded as character strings. Each voter is assigned a unique voter ID, which serves as a point of reference between the two files. Broadly speaking, data in this file can be divided between three categories: first, personal identification information like address, ZIP code, or phone number; second, demographic information like age and gender; third, information pertinent to elections administration like congressional district, local elections for which the individual should receive a ballot, voter ID, and party registration. I will further elaborate on relevant variables in the wrangling section.

In the Voter History files, the unit of observation here is a single ballot cast, and all variables are initially coded as character strings. This means that for each voter registered–and so included in the VRF–the history file should contain an observation for each time they voted. This file includes two types of data: first, identifiers for the election like county, date, description, and type; second, identifiers for the individual vote including voter ID and voting method.

3.3 Wrangling the Data

The process of "wrangling" refers to manipulating the data into a form that can then be used for graphing, exploratory data analysis, modelling, or presentation. In this case, wrangling also included aggregating data across multiple sources and datasets. For this purpose, I made heavy use of the tidyverse R package, and in particular the dplyr package. In this section I will go through some of the key problems encountered during the wrangling of these data, and then discuss the final form each variable takes.

3.3.1 Initial Problems with the 2017 Voter File and Solution

The first major issue I encountered–which merits discussion in its own section–derives from the aforementioned fact that the voter records I had access to are "snapshots". What this means, is that for each person in each year of voter registration files, I will have their corresponding history files for all ballots they have cast in Colorado, but not their own history of registration and migration. If, say, a voter moved from Boulder County to Summit County, I would have their votes in Boulder County show up in the voter history file, but them being registered in Summit. If you recall the turnout calculations specified earlier on, this implies an overestimation when looking back at elections that happened some time before the date of the "snapshot". Additionally, "snapshots" of current voter files do not reflect voters dropping off the rolls for whatever reason–death, moving out of the state, long term inactivity, non-confirmable personal data etc. Since for these voters the history files would also not be included, the issue created is less one of overestimation of turnout like before, but just the inclusion of additional room for error that is created when subtracting one from the denominator and enumerator of turnout.

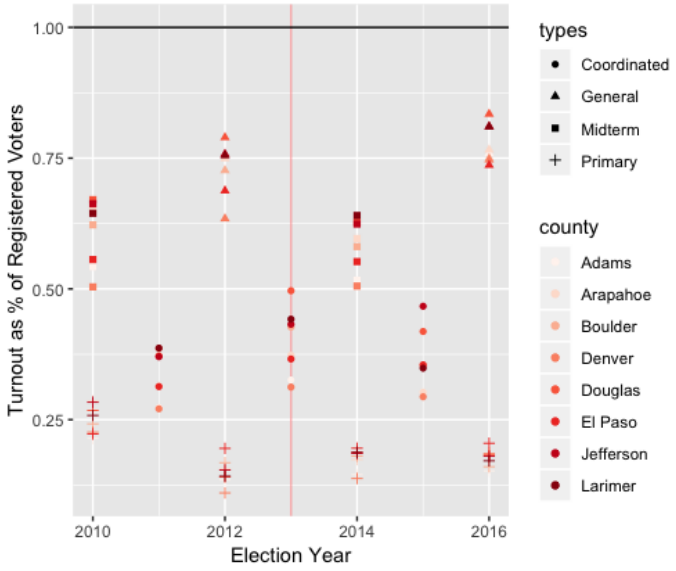

Figure 3.4: Turnout plot for eight largest Colorado counties, 2012-2016

This was a significant problem from the beginning of this thesis, since I started out with only one "snapshot" from 2017. After going through turnout calculations, a significant majority of counties appeared to have turnout exceeding 100%, particularly for years between 2000 and 2012. This was, to put it mildly, concerning. With the help of my advisers, I was able to procure similar "snapshots" for each year between 2012-2016. After similar calculations, I returned figure 3.4 for the eight most populous counties as described above, including different shapes for election type, colors for county, and a vertical line at 2013 to signify the latest major change in how Colorado administers elections.

To also further illustrate the in-county migration and dropped voter problem, I created a graph that includes logged total counts of registered voters calculated using the 2017 and the 2012-2016 files. The plot also includes a line at y=x. If in-Colorado migration and dropped voters are not an issue, most points on this graph should be at this line.

Two things should be clear from figure 3.5. First, there is significant deviation between the counts using just the 2017 file and all files across years. Specifically, the 2017 count consistently underestimates the total amount of registered voters–this is shown by the red linear model smoothing line. This consistent difference confirms the hypothesis that there is a substantial benefit to using "snapshots" for multiple years. Second, counts get more accurate the closer to 2017 we get. This should be even more apparent in figure 3.6, which limits the scale to only some high registration counties, and adds a shape indicator for county.

Here the structure of the data becomes clear: for each county, there are a series of almost vertically distributed points, which get closer to the y = x line the closer the counts get to 2017. Through this series of tests, it became clear that using multiple years of data was necessary in order to conduct an accurate test of my hypotheses.

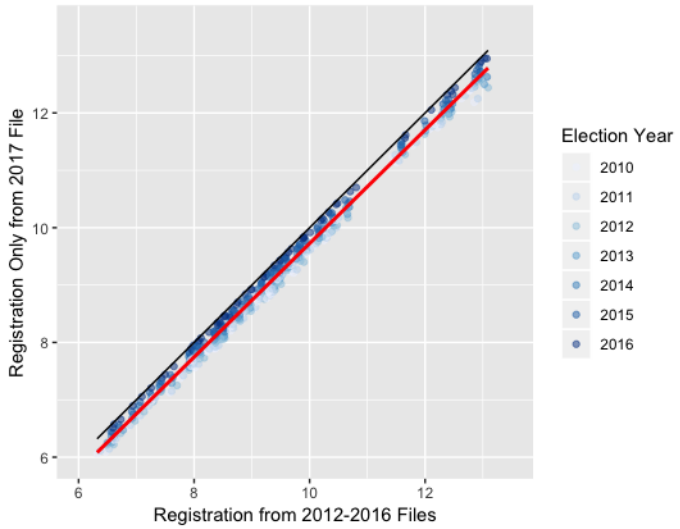

Figure 3.5: Comparison of registration count methods

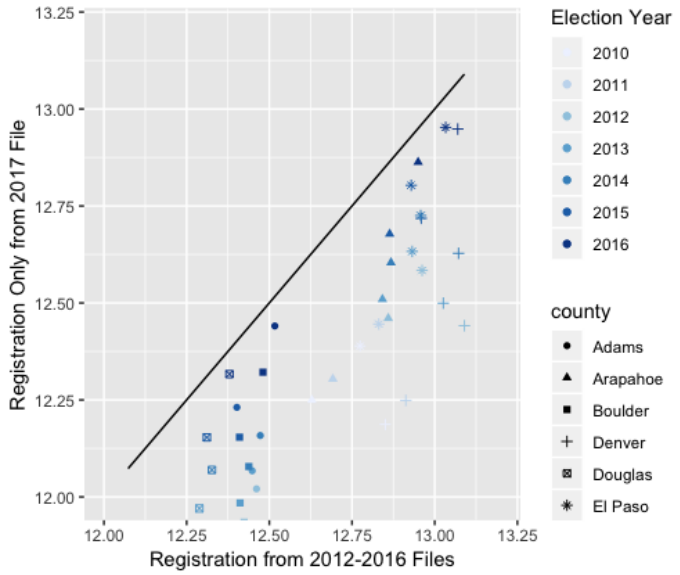

Figure 3.6: Comparison of registration count methods only for a few counties, 2012-2016

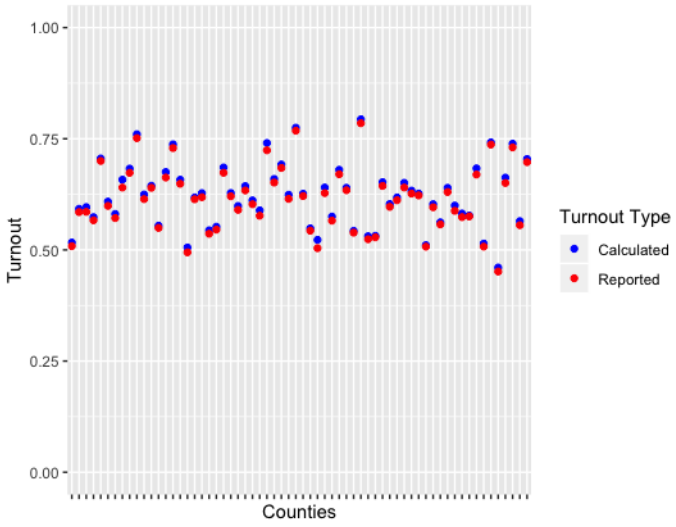

Figure 3.7: Comparison of reported and calculated turnout for 2014 midterms across county

My selection was later vindicated, when looking at comparisons between reported rates of turnout[2] and turnout calculated through my dataset for the 2014 midterm election (see fig. 3.7).

The differences are insignificant. They exist because of "noise" added on because of errors in the data, misreporting, private voter registration files, voters dropped before the "snapshot" occurred, and other similar factors.

3.3.2 Other Wrangling Issues Faced

Suffice to say, wrangling data was the majority of the work that went into this thesis. Doing a full account would probably read like the world's most cliche crime novel: a series of elusive final datasets, a plucky yet occasionally naive young detective, two wisened mentors, clues, dead ends, frustration, compromise, and. . . spreadsheets. I will spare the reader the whole story, but I will include a non-comprehensive list of some of the difficulties associated with wrangling voter files, as it was a crucial part of the learning process I underwent while doing my research.

Missing Values: The decision on how to deal with missing values–or NAs–in a dataset is a lot more important than it may initially seem. A first, intuitive reaction might be to just disregard them; however this works under the assumption that there is no structure inherent to why these data are missing! To give just two examples, in the data I have collected, the PARTY value for the 2015 voter registration file is missing. If I excluded all observations with missing PARTY values, I would be excluding a fifth of my data. Missing values were also present in the VOTING_METHOD variable of the voter history files. While this may have seemed troubling, after closer examination it was revealed that the vast majority of such missing values was concentrated in

[2]Turnout is calculated over all registered voters

Jefferson County, and in elections prior to 2002. Therefore, these observations could be ignored, since they played no role in my final dataset. The conclusion should be that choices made on exclusion, inclusion, or estimation of missing data are very important, and should be taken with much care and consideration for the underlying structure of the data.

Data Input Errors: Is "Greece" a legitimate voting method? Probably not. However, "Greece" did show up as a value in the VOTING_METHOD variable for my 2012 voter history file snapshot. This may have occurred for a series of reasons, like data reading issues–the data I acquired had changed hands some times, and also changed platforms between STATA and R–or issues at time of input–each county counts votes individually, and *then* the state aggregates the data–, or some bug in my code. Having adequately checked for the later of these reasons, I treated all values that seemed more likely than not to be errors as NAs. There were not many of these–less than .001% of my data–but they were a hassle to find, analyze, and then recode into some useful value.

Data Size: Nothing to write home about here, just an observation that multiple voter registration files can be *huge*, which puts considerable strain on a computer's processing power. This means that wrangling has to comprise of a series of careful, deliberate moves. Brute force should be discouraged, as a dead end means several hours of melodic computer fan panic.

Joining, Merging, Spreading, and the Multiplicity of Levels: For the data to end up in any functional shape, it eventually becomes necessary to start joining datasets. Thankfully, a clear division of modelling tasks between county and individual level models means that joining on COUNTY or VOTER_ID is ideal, and fairly straightforward. As will become clear in later sections, I also had to consider the variety of different units of observation, specifically: county, individual, ballot, election, county-by-election.

3.3.3 Final Variable Specifications

After the conclusion of the wrangling process, the resulting dataset included a series of discrete and continuous variables. I will briefly outline them here, along with their range and values.

- VOTER_ID: Discrete variable, unique value given to each individual voter. Useful for merging.
- COUNTY: Discrete variable, the 64 counties of Colorado.
- REGISTRATION_DATE: Discrete variable, date of registration for each registrant. Useful to get total registrants on election day.
- TURNOUT: Continuous variable, in the range [0,1]. The response variable for my county-level models.
- ELECTION_TYPE: Discrete variable, the four types of elections: Primary, Coordinated, Midterm, Presidential.
- ELECTION_DATE: Discrete variable, self-explanatory.

- VBM_PCT: Continuous variable, in the range [0,1]. This is the focus of my analysis, as it counts the percentage of total ballots that were mail ballots.
- PCT_WHITE: Continuous variable, in the range [0,1]. Percentage of white residents per county.
- PCT_URBAN: Continuous variable, in the range [0,1]. Percentage of urban residents per county.
- PARTY: Discrete variable. For each voter, the party they are registered with. Can be: Republican, Democrat, Other, or Unaffiliated.
- GENDER: Discrete binary variable, Male or Female.
- AGE: The age of the individual registrant.
- VOTING_METHOD: The method used by an individual voter to cast their ballot. Is coded as either VBM or In Person, according to Table 3.4:

Table 3.4: Voting method designation table

Voting Method	Description of Method	Final Designation
Absentee Carry	Voters who carried an absentee ballot with them from an early voting location	VBM
Absentee Mail	Voters who were sent an absentee ballot, and mailed it in	VBM
Early Voting	Voters who physically went to an Early Voting location and voted	In Person
In Person	Voters who physically went to a polling place and voted on paper	In Person
Mail Ballot	Vote By Mail	VBM
Polling Place	Traditional polling place voting, discontinued in 2013	In Person
Vote Center	Voters who cast their ballots at Vote Centers	In Person

Chapter 4

Model Specification and Results

In this chapter I do a step-by-step construction and fitting of a series of models. I begin with a thorough analysis of my notation, and specification of models. I then extract results from the models that best fit the data, and draw inferences on my hypotheses. I will start with a disclaimer: why the data available to me is not necessarily enough to get the necessary causal leverage for significant results. This is particularly true of individual level models. This disclaimer should be considered a large part of the analytical results of my thesis; given several months of exploratory data analysis, such explanations should serve future research into Colorado voter files. I will then proceed to construct some actual models–disclaimer notwithstanding–starting with county and proceeding to individual level modelling. Some of these models may be speculative, as given the initial disclaimer they require more data, or more processing power to actually run. I will still include them as they may be used for future research.

4.1 Salt

The following models should be taken with a grain of salt because of a series of reasons. In this section, I will tackle these reasons one by one and then analyze what steps could be made to compensate.

4.1.1 Causal Leverage

Causal leverage usually means having enough data to draw significant conclusions about correlation of variables, or in the best cases make safe causal inferences. Data that presents causal leverage should have certain characteristics. It should, first, be extensive enough. By this I mean that the raw number of observations should be as high as possible, and in the best cases significantly larger than the set of predictors that are present. This not only guarantees no issues with model matrices when running statistical models, but also that there will be enough data-per-variable to draw conclusions. Second, the data should be varied. To put it very simply, it's not enough to have hundreds of thousands of observations if they are all similar to eachother. If, for example, my data included a thousand people in Jefferson county,

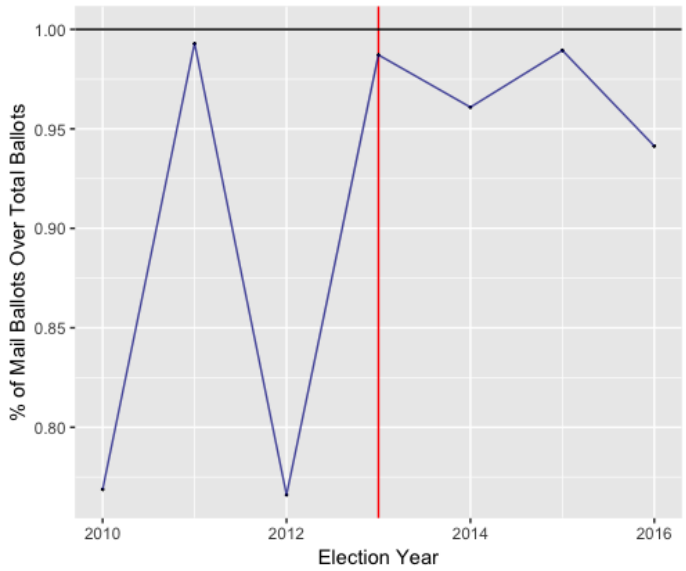

Figure 4.1: Percentage of mail ballots over total ballots by year

and 63 in all other counties of Colorado combined–one in each remaining county–, then I would not be able to leverage my data to draw conclusions on county-level effects.

As previously stated, I have registration files going back to 2012. From these files, I have extracted data for elections going back to 2010.[1] In order to make inferences on VBM and turnout effects, given the previous criteria for causal leverage, I had to have extensive and varied data. I have extensive data–over 35 million observations at the individual level–but the data substantially lacks variance in voting method. Put simply, most people in Colorado from 2010 onward either did not vote at all, or voted by mail. If you recall the changes in Colorado election law, in 2008 counties were allowed to conduct all mail elections, and no-excuse permanent absentee voting was implemented state-wide; then in 2013 Colorado transitioned to full VBM for all elections. This means that few people were still using traditional polling places or vote centers to cast their ballots. Figure 4.1 shows how, after 2013, and even before that in 2011–the coordinated, local election for which mail ballots were more convenient for counties–over 95% of ballots cast were mail ballots. Only in the general elections of 2010 and 2012 there is some variance, but mail ballots account for well over two thirds of total votes.

This issue is not completely fatal for my county level models. There is still variance between counties that have 100% mail ballots and those that are around the 75-85% margin. For individual level models–where I am estimating voting probability–VBM will be an almost perfect predictor for voting, and therefore will not present me with any substantial analytical result on how it affects voting probability. There are some ways to compensate for this issue, which I outline; due to time or data constraints,

[1]See section 3.3.1; I extracted data limited to this time period to avoid accuracy issues with migration and removal of inactive/unavailable voters

not all of these will be implemented in this thesis:

- *More (Diverse) Data*:
- *Localized, Natural Experiment Studies*:
- *Synthetic Control Group*:

4.1.2 Lack of Individual Data

4.1.3 Processing Power

4.2 Models

4.2.1 Variable Specification

I will not go through each individual variable in this section, but will briefly describe my procedure on notation for the following models. I will include more comments whenever they seem necessary under each model. In this thesis I include predictors on a series of variables that can be divided into five categories based on unit of observation: county, election, individual, local result, and ballot. The last two are functions of other units: local result units are equal to the product of elections and counties, while ballot units are equal to the number of unique individuals multiplied by the number of elections each of them was registered in. For notation, I follow this set of rules:

1. If the variable is a response, it is coded y.
2. If the variable is a predictor, it is coded according to Table 4.1.
3. The variable's superscript will provide information on what it represents, else it will be explained.
4. All variables represent a single value of that variable unless stated otherwise.
5. Unit of observation will also be specified in subscript, according to the indices described in Table 4.1. These indices are also used in sum notation.
6. All Greek characters represent coefficients to be calculated.
7. By $k[j]$ I represent the k-value of the j-observation. In this case, this would be the county that an individual is registered in.
8. Note that for Local Result level variables, I use k, l as an indice. This is because there are very few variables at this level, it is a direct multiplicative product of two other units, and this notation avoids confusion with even more indice types.

Table 4.1: Variable names and indices per unit of observation

Units	Variable	Index
Ballot	u	i
Individual	z	j
County	x	k

Units	Variable	Index
Election	w	l
Local Result	v	k,l
General Index	-	i'

4.3 County Level Models

4.3.1 Specifications

In this section I will go through a step-by step creation of models at the county level. County level models use a series of variables at the election, county, and local result levels. The response variable is always turnout as a local result. If this model is considered at its most basic, it could be thought of as an assignment of voting tendencies across counties; each county independent of election has a unique range of turnout results. In this way it is possible to build a naive, baseline model of turnout as follows:

$$y_{k,l}^{turnout} = \beta_0 + (\sum_{k=1}^{64} \beta_k x_k^{county}),$$

where x_k^{county} is a series of 64 dummy variables for each county of Colorado. Here differences between elections come from normally distributed error terms, rather than predictors. I name this *Model 1*, and it does not reflect the data particularly well. First off, this model includes the assumption that counties are independent of one another, which is probably false; just consider that these counties are areas of the same state, in the same country, with populations moving between them at regular intervals, and many of them covering the same metropolitan area or congressional district. Additionally, this model cannot fully calculate relevant coefficients, since a number of counties can be represented as perfect linear functions of the other variables. This means they will be dropped by *R* when the model is called in the `lm()` function.

A way to fix both these issues is to use a multilevel model with mixed effects for county. By constraining coefficients at the county level to a set distribution, this model does away with the assumption of independence. The other county level predictors help to explain some of the unexplained group level variation, which reduces the standard deviation of county coefficients and helping provide more exact estimates (Gelman & Hill, 2006). I call this *Model 2*, which can be written as:

$$y_{k,l}^{turnout} = a_k + \beta_1 x_k^{\%white} + \beta_2 x_k^{\%urban},$$

$$a_k \sim N(\gamma_0, \sigma_\alpha^2)$$

This model provides a more reasonable set of estimates for each county, but still fails at providing any sort of guess as to secular trends, time-specific effects, election type effects, or mail voting–the variable of interest. I will amend this by adding a set

of variables at the election and local result levels: election type and an interaction term between election type and mail voting. This variable should reflect whether turnout effects of mail voting are more pronounced in a specific type of election. I call this *Model 3* and it can be specified as follows:

$$y_{k,l}^{turnout} = a_k + \beta_1 x_k^{\%white} + \beta_2 x_k^{\%urban} + (\sum_{i'=1}^{4} \beta_{i'+3} w_{i'}^{electiontype}) * (\beta_3 v_{k,l}^{\%mail\ vote} + 1),$$

$$a_k \sim N(\gamma_0, \sigma_\alpha^2)$$

where $w_{i'}^{electiontype}$ is a series of four dummy variables for each type of election (General, Primary, Coordinated, Midterm). This model reflects nearly all the information I have available, apart from election date. For the incorporation of election dates there are two possible alternatives. First, I can simply add a dummy variable for each year. This would assume independence between each year, as it would specify different, independent "slopes" for the seven years I have data for–this is like calculating seven different models, one for each year. This is not particularly elegant as a solution nor does it reflect the fact that years actually are interconnected; of course there can be massive shifts in national or regional political climates, but those shifts happened *from some baseline*, which is reflected in previous years.

These elections can be thought of as systems for which prior condition affects future outcomes, and therefore time cannot be modeled as a series of independent effects. The solution here is adding a spline function for time, using a general additive multilevel model. The most commonly used spline function, and the default in the `gamm4` *R* package is a thin plate regression spline, which I also use here (S. N. Wood, 2006). More on the subject of splines can be found in the Wood (2006) textbook. The model, which I call *Model 4* can be written as follows:

$$y_{k,l}^{turnout} = a_k + \beta_1 x_k^{\%white} + \beta_2 x_k^{\%urban} + (\sum_{i'=1}^{4} \beta_{i'+3} w_{i'}^{electiontype}) * (\beta_3 v_{k,l}^{\%mailvote} + 1) + s(w_l^{year}),$$

$$a_k \sim N(\gamma_0, \sigma_\alpha^2)$$

where $s()$ is a thin plate spline function with seven knots–equal to the number of years.[2] A summary of these four models is provided in the following table:

Table 4.2: County level model descriptions

Model No	Model Description
Model 1	Naive model with only county specific effects
Model 2	Multilevel model; added county level predictors

[2]I used the `gam.check()` function that is present in the `mgcv` *R* package, whose call determined that the number of knots here may be too low. However, given the data available to me, I was limited to the inclusion of seven years and as such cannot increase the number of knots any further.

Model No	Model Description
Model 3	Multilevel model; added VBM, interaction terms, and election fixed effects
Model 4	Multilevel General Additive model; added spline function for election year

4.3.2 Results

Calling model one results in the following:

```
md_1 <- lm(data = model_dt, turnout ~ pct_white + pct_urban + county)

summary(md_1)
```

```
Call:
lm(formula = turnout ~ pct_white + pct_urban + county, data = model_dt)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38095 -0.15980 -0.03896  0.17156  0.56682

Coefficients: (2 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.684609   0.930859   0.735   0.4623
pct_white         -0.091375   1.004528  -0.091   0.9276
pct_urban         -0.277328   0.393461  -0.705   0.4812
countyAlamosa     -0.008657   0.205569  -0.042   0.9664
countyArapahoe     0.059868   0.108177   0.553   0.5802
countyArchuleta   -0.074504   0.080693  -0.923   0.3562
countyBaca        -0.038293   0.125250  -0.306   0.7599
countyBent         0.031705   0.125245   0.253   0.8002
countyBoulder      0.058465   0.221585   0.264   0.7920
countyBroomfield   0.118109   0.252452   0.468   0.6401
countyChaffee      0.096121   0.189179   0.508   0.6116
countyCheyenne    -0.004180   0.124061  -0.034   0.9731
countyClear Creek -0.158121   0.117195  -1.349   0.1777
countyConejos     -0.147084   0.518922  -0.283   0.7769
countyCostilla    -0.140564   0.626891  -0.224   0.8227
countyCrowley     -0.114716   0.363623  -0.315   0.7525
countyCuster       0.001746   0.117203   0.015   0.9881
countyDelta       -0.024595   0.093877  -0.262   0.7934
countyDenver      -0.002660   0.086058  -0.031   0.9753
countyDolores     -0.115033   0.117797  -0.977   0.3292
```

```
countyDouglas      0.095437   0.272803   0.350   0.7266
countyEagle       -0.029321   0.084352  -0.348   0.7283
countyEl Paso      0.032358   0.153918   0.210   0.8336
countyElbert      -0.098869   0.117751  -0.840   0.4014
countyFremont      0.039704   0.169325   0.234   0.8147
countyGarfield     0.002705   0.083721   0.032   0.9742
countyGilpin      -0.195298   0.117816  -1.658   0.0979 .
countyGrand       -0.096441   0.107068  -0.901   0.3681
countyGunnison    -0.091462   0.146485  -0.624   0.5326
countyHinsdale     0.054785   0.117729   0.465   0.6418
countyHuerfano     0.018049   0.161901   0.111   0.9113
countyJackson     -0.034279   0.126263  -0.271   0.7861
countyJefferson    0.094609   0.234286   0.404   0.6865
countyKiowa        0.007118   0.117766   0.060   0.9518
countyKit Carson   0.042564   0.080601   0.528   0.5976
countyLa Plata    -0.108083   0.085798  -1.260   0.2082
countyLake        -0.044992   0.110657  -0.407   0.6844
countyLarimer      0.072687   0.260269   0.279   0.7801
countyLas Animas  -0.006712   0.177205  -0.038   0.9698
countyLincoln     -0.071115   0.172840  -0.411   0.6809
countyLogan        0.091242   0.140329   0.650   0.5158
countyMesa         0.042887   0.243655   0.176   0.8603
countyMineral      0.016145   0.121265   0.133   0.8941
countyMoffat       0.007592   0.187606   0.040   0.9677
countyMontezuma   -0.114966   0.099263  -1.158   0.2472
countyMontrose     0.016583   0.091876   0.180   0.8568
countyMorgan       0.016310   0.089452   0.182   0.8554
countyOtero       -0.001323   0.134714  -0.010   0.9922
countyOuray       -0.116254   0.117889  -0.986   0.3244
countyPark        -0.122114   0.117321  -1.041   0.2983
countyPhillips    -0.119730   0.173567  -0.690   0.4906
countyPitkin      -0.065013   0.178769  -0.364   0.7162
countyProwers     -0.008041   0.096114  -0.084   0.9334
countyPueblo       0.016798   0.096983   0.173   0.8625
countyRio Blanco  -0.075465   0.130767  -0.577   0.5641
countyRio Grande  -0.066331   0.247587  -0.268   0.7889
countyRoutt       -0.028231   0.201408  -0.140   0.8886
countySaguache    -0.195286   0.377335  -0.518   0.6050
countySan Juan    -0.163076   0.136682  -1.193   0.2333
countySan Miguel  -0.217148   0.122666  -1.770   0.0772 .
countySedgwick    -0.014318   0.134360  -0.107   0.9152
countySummit      -0.083251   0.215208  -0.387   0.6990
countyTeller      -0.061382   0.147748  -0.415   0.6780
countyWashington   0.009425   0.120239   0.078   0.9375
countyWeld               NA         NA      NA       NA
```

```
countyYuma                NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1994 on 640 degrees of freedom
Multiple R-squared:  0.1251,    Adjusted R-squared:  0.03899
F-statistic: 1.453 on 63 and 640 DF,  p-value: 0.01564
```

```
#If run this shows perfect linear relationship
#alias(md_1)

#plot(md_1)
```

Calling Model two has the following result:

```
md_2 <- lmer(data = model_dt, turnout ~ pct_white + pct_urban + (1|county),
             REML = F)

arm::display(md_2)
```

```
lmer(formula = turnout ~ pct_white + pct_urban + (1 | county),
    data = model_dt, REML = F)
            coef.est coef.se
(Intercept)  0.49     0.05
pct_white    0.03     0.05
pct_urban   -0.12     0.02

Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.00
 Residual             0.20
---
number of obs: 704, groups: county, 64
AIC = -270.8, DIC = -280.8
deviance = -280.8
```

```
#Run for specific county coefs
#ranef(md_2)

fixef(md_2)
```

```
(Intercept)   pct_white   pct_urban
 0.49202848  0.03364891 -0.11827707
```

```
#plot(md_2)

#qqnorm(residuals(md_2))
```

Calling Model three:

```
md_3 <- lmer(data = model_dt, turnout ~ 1 + types + pct_vbm +
               pct_urban + pct_white + pct_vbm:types + (1|county),
             REML = F)

arm::display(md_3)
```

```
lmer(formula = turnout ~ 1 + types + pct_vbm + pct_urban + pct_white +
    pct_vbm:types + (1 | county), data = model_dt, REML = F)
                    coef.est coef.se
(Intercept)          0.45     0.08
typesGeneral         0.19     0.07
typesMidterm         0.25     0.07
typesPrimary        -0.07     0.07
pct_vbm              0.00     0.07
pct_urban           -0.12     0.02
pct_white            0.03     0.05
typesGeneral:pct_vbm  0.15     0.07
typesMidterm:pct_vbm -0.06     0.07
typesPrimary:pct_vbm -0.09     0.07

Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.05
 Residual             0.06
---
number of obs: 704, groups: county, 64
AIC = -1779.2, DIC = -1803.2
deviance = -1803.2
```

```
#Run for county coefs
#ranef(md_3)

fixef(md_3)
```

```
        (Intercept)         typesGeneral         typesMidterm
          0.454642317          0.190033387          0.252085742
       typesPrimary              pct_vbm            pct_urban
         -0.070582336         -0.001196972         -0.116841416
```

```
          pct_white typesGeneral:pct_vbm typesMidterm:pct_vbm
        0.032823780          0.151950039         -0.057236047
typesPrimary:pct_vbm
       -0.088157086
```

```
#plot(md_3)

#qqnorm(residuals(md_3))

anova(md_2, md_3)
```

```
Data: model_dt
Models:
md_2: turnout ~ pct_white + pct_urban + (1 | county)
md_3: turnout ~ 1 + types + pct_vbm + pct_urban + pct_white + pct_vbm:types +
md_3:     (1 | county)
     Df      AIC      BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
md_2  5  -270.82  -248.03 140.41  -280.82
md_3 12 -1779.20 -1724.52 901.60 -1803.20 1522.4      7  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calling Model 4:

```
model_dt$dates <-  as.integer(model_dt$dates)

md_4 <- gamm4(turnout ~ 1 + types +
                pct_urban + pct_white + pct_vbm*types + s(dates, k = 7),
              random =~ (1|county),
              data = model_dt)

summary(md_4$mer)
```

```
Linear mixed model fit by REML ['lmerMod']

REML criterion at convergence: -1899.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1206 -0.6102 -0.1110  0.5546  4.9920

Random effects:
 Groups   Name        Variance Std.Dev.
 county   (Intercept) 0.002962 0.05442
```

```
 Xr       s(dates)    0.967661 0.98370
 Residual             0.002788 0.05281
Number of obs: 704, groups:  county, 64; Xr, 5

Fixed effects:
                       Estimate Std. Error t value
X(Intercept)           0.469534   0.072036   6.518
XtypesGeneral          0.254063   0.064603   3.933
XtypesMidterm          0.070291   0.062977   1.116
XtypesPrimary         -0.170327   0.061898  -2.752
Xpct_urban            -0.119413   0.020723  -5.762
Xpct_white             0.031336   0.050401   0.622
Xpct_vbm               0.002371   0.058353   0.041
XtypesGeneral:pct_vbm  0.085084   0.067613   1.258
XtypesMidterm:pct_vbm  0.106871   0.064296   1.662
XtypesPrimary:pct_vbm -0.005732   0.061585  -0.093
Xs(dates)Fx1          -0.113090   0.019823  -5.705

Correlation of Fixed Effects:
            X(Int) XtypsG XtypsM XtypsP Xpct_r Xpct_w Xpct_v XtyG:_ XtyM:_
XtypesGenrl -0.715
XtypesMdtrm -0.741  0.822
XtypesPrmry -0.736  0.833  0.882
Xpct_urban  -0.292 -0.001  0.000  0.005
Xpct_white  -0.571  0.001  0.000 -0.002  0.336
Xpct_vbm    -0.792  0.864  0.887  0.883 -0.008 -0.003
XtypsGnrl:_  0.661 -0.967 -0.747 -0.764  0.000 -0.002 -0.836
XtypsMdtr:_  0.705 -0.779 -0.968 -0.833 -0.001 -0.001 -0.880  0.751
XtypsPrmr:_  0.719 -0.808 -0.846 -0.972 -0.005  0.002 -0.903  0.789  0.846
Xs(dats)Fx1 -0.015  0.116  0.138  0.107  0.011  0.004  0.013 -0.156 -0.146
            XtyP:_
XtypesGenrl
XtypesMdtrm
XtypesPrmry
Xpct_urban
Xpct_white
Xpct_vbm
XtypsGnrl:_
XtypsMdtr:_
XtypsPrmr:_
Xs(dats)Fx1 -0.112
```

```
#plot(fitted(md_4$mer), residuals(md_4$mer))
```

```
#qqnorm(residuals(md_4$mer))
```

4.4 Individual Level Models

4.4.1 Specifications

For the rest of this write-up, assume the following:

$$y_i \sim \text{Bernoulli}(p\_vote)$$

Where $y_i \in \{0, 1\}$ is the probability that the i-th ballot was completed.

In this section I do not linearly add to the model until it reaches a final stage. The reasoning here is that there is no exact linear path to follow; there is an overarching unit of observation–the ballot–and all the rest are dependent between each other. For instance, adding a variable for Party at the ballot level would not significantly change the way I later add percentage of white residents at the county level. Therefore, the way I proceed is the following: I "build" the models step by step and separatelly for each group of variables (grouping by unit of observation). Then I present one example of what a model using two of these initial "building blocks" would look like. Since this is fairly generalizable, I then proceed directly to the full model which inlcudes all different variables.

If receiving a ballot with no information, I would predict that the probability that an additional ballot was a vote in favor would be equal to turnout, as calculated through all other ballots. Therefore:

$$\hat{p\_vote}_i = \frac{\#\text{votes cast}}{\#\text{ballots}}$$

4.4.2 Estimation with only one type of data

There are four levels of data I will go through here: County, Election, Person, and Ballot.

County Level

Assume that the ballot I am trying to assess completion for has the name of the county it is from written on it. There are two ways I can think of for predicting p_vote. First, assume that each different county has a different, independent p_vote. Therefore, in model-lingo this would look like:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_{k,i}\beta_k)$$

Where k counts over the 64 counties of Colorado, and x_k is an indicator variable for each county. If I, quite reasonably, throw away the assumption of independance–these

counties are, after all, in the same state and the same country–I could also fit a mixed effects model as such:

$$p\_\hat{v}ote \sim \text{logit}^{-1}(a_{k[i]}),$$
$$a_k \sim \text{N}(\gamma_0, \sigma_\alpha^2)$$

Where $\alpha_{k[i]}$ varies by county, constrained by its standard deviation and γ_0, an intercept coefficient. Let's say now that along with the one ballot, I was given a short list of $n^{\text{county vars}}$ other county-level variables, be they discrete, continuous, or indicators. The two models would then look like:

$$p\_\hat{v}ote_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_k\beta_k + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\beta_{i'+64})$$

Where $x_{k[i],l}$ is the k-th value of the i'-th variable. If, as before, I do not assume independance, the model can be written as:

$$p\_\hat{v}ote \sim \text{logit}^{-1}(a_{k[i]}),$$
$$a_k \sim \text{N}(\gamma_0 + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\gamma_{i'}, \sigma_\alpha^2)$$

In the case of my specific data, for the time being I have county-level data for white population and urban population, so $n^{\text{county vars}} = 2$.

Individual Level

Assuming that I know the voter ID of the individual that cast their ballot, I can treat this piece of information in about the same way that I did for county as described above. This means that the following is mostly an exercise in maintaining notation constant. For these purposes, let n^{ID} be the number of total unique voter IDs–individuals–that I have data on, and j an indice that sums over all individuals. Also let z_j be an indicator variable for each individual. Then:

$$p\_\hat{v}ote_i \sim \text{logit}^{-1}(\sum_{j=1}^{n^{ID}} z_j\beta_j)$$

And the second model, not assuming independence, would be:

$$p\_\hat{v}ote \sim \text{logit}^{-1}(\delta_{j[i]}),$$
$$\delta_j \sim \text{N}(\zeta_0, \sigma_\delta^2)$$

Again, in a similar way to county level data, there are variables at an individual level, thus making it relativelly easy to build further models. Let's say now that along with the one ballot, I was given a short list of $n^{\text{indiv vars}}$ other individual-level variables, be they discrete, continuous, or indicators. The two models would then look like:

$$p\_\hat{v}ote_i \sim \text{logit}^{-1}(\sum_{j=1}^{n^{ID}} z_j\beta_j + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\beta_{i'+n^{ID}})$$

Where $z_{j[i],l}$ is the j-th value of the i'-th variable. If, as before, I do not assume independance, the model can be written as:

$$p\_\hat{v}ote \sim \text{logit}^{-1}(\delta_{j[i]}),$$
$$\delta_j \sim \text{N}(\zeta_0 + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\delta_{i'}, \sigma_\delta^2)$$

In the case of my specific data, for the time being I have individual-level data for gender, so $n^{\text{indiv vars}} = 1$.

Election Level

Again as previously, four models come from including election level data. The first two are assuming I only knew what specific election the ballot comes from. Let $w_{i'}$ be an indicator variable for each election and n^{elect} the number of elections. The model assuming independence, with $w_{i'}$ being indicator variables for each election, is:

$$p\_\hat{v}ote_i \sim \text{logit}^{-1}(\sum_{l=1}^{n^{elect}} w_l\beta_l)$$

Again, as previously, it would be safe to assume that each election is not held in a vacuum. Adding mixed effects this model would be:

$$p\_\hat{v}ote \sim \text{logit}^{-1}(\eta_{l[i]}),$$
$$\eta_l \sim \text{N}(\nu_0, \sigma_\nu^2)$$

Again, in a similar way to county and individual level data, I add in variables at an election level. Let's say now that along with the one ballot, I was given a short list of $n^{\text{election vars}}$ other election-level variables, be they discrete, continuous, or indicators. The two models would then look like:

$$p\_\hat{v}ote_i \sim \text{logit}^{-1}(\sum_{l=1}^{n^{elect}} w_l\beta_l + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\beta_{i'+n^{elect}})$$

Where $w_{l[i],i'}$ is the l-th value of the i'-th variable.

Assuming independence:

$$p\_\hat{v}ote \sim \text{logit}^{-1}(\eta_{l[i]}),$$
$$\eta_l \sim \text{N}(\nu_0 + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\nu_{i'}, \sigma_\nu^2)$$

For the time being I have two different variables that describe individual elections: date and type. Note that the above models may not be the best way to describe dates!

An alternative could be fitting a glm, with some smoothing spline function for year. As for type, this would include four distinct indicators; one for each election type.

Ballot Level

In this section I assume that the ballot has some key features written on it, like the voting method, age, or party registration of the person that filled it out. A mixed effects model here would make no sense, since all the data is at the same unit of observation. Therefore, when adding ballot level variables, the model would look like:

$$p\_\hat{v}ote_i \sim \text{logit}^{-1}(\beta_0 + \sum_{i'=1}^{n^{\text{ballot vars}}} u_{i,i'}\beta_{l'})$$

Where $u_{i,i'}$ is the i-th value of the i'-th variable, and $n^{\text{ballot vars}}$ is the number of ballot level variables. For now, I have data on voting method, age, and party. Voting method is coded as a binary variable with value one if the method was a Mail Vote. Party includes four distinct indicators for REP, DEM, Other, and Unaffiliated. Age is tricky; for now the options would be: straight up inclusion as an integer, inclusion as a cubic polynomial, unclusion as a 2nd degree polynomial, inclusion in some form of spline function.

4.4.3 Estimation with two types of data

After the work of setting up the four models at four different levels of observation, combining them in twos should be failry straightforward. To avoid being nedlessly cumulative, I will pursue this combination for County and Individual level only–instead of the six different possible combinations.

With the assumption that both counties and individuals are independent of one another, I proceed to the first type of model:

$$p\_\hat{v}ote_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_k\beta_k + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\beta_{i'+64} + \sum_{j=1}^{n^{ID}} z_j\beta_{j+n^{\text{county vars}}+64} + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\beta_{i'+n^{ID}+n^{\text{county va}}}$$

This is large and clunky. It includes variables as described above: indicators for each county and individual, and all individual or county-level variables. For the corresponding mixed-effects model, I assume the tree-like structure we discussed on Monday. The hierarchy has two "levels", with the second level consisting of two different regressions:

$$p\_\hat{v}ote \sim \text{logit}^{-1}(\delta_{j[i]} + a_{k[i]}),$$
$$a_k \sim \text{N}(\gamma_0 + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\gamma_{i'}, \sigma_\alpha^2)$$
$$\delta_j \sim \text{N}(\zeta_0 + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\delta_{i'}, \sigma_\delta^2)$$

4.4.4 Estimation with the full dataset

I now proceed to include variables from all units of observation into one model. The first model, assuming independence, is:

$$\hat{p\_vote}_i \sim \text{logit}^{-1}(\sum_{k=1}^{64} x_k\beta_* + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\beta_* + \sum_{j=1}^{n^{ID}} z_j\beta_* + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\beta_* + \sum_{l=1}^{n^{elect}} w_l\beta_* + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\beta_* + \sum_{i'=1}^{n^{\text{ballot vars}}} u_{i,i'}\beta_*)$$

You will notice that I have ommited the subscript for all beta coefficients. This is because after two or three parameters, this becomes very, very large. I think it's reasonable to assume increasing indexes for different beta coefficients from left to right in this expression.

The mixed effects model will again operate on two "levels" of hierarchy, but the second level will now include three distinct regressions. Caveats for variables like age and date should be noted from previous sections.

$$\hat{p\_vote} \sim \text{logit}^{-1}(\sum_{i'=1}^{n^{\text{ballot vars}}} u_{i,i'}\beta_{l'} + \delta_{j[i]} + a_{k[i]} + \eta_{l[i]}),$$

$$a_k \sim \text{N}(\gamma_0 + \sum_{i'=1}^{n^{\text{county vars}}} x_{k[i],i'}\gamma_{i'}, \sigma_\alpha^2)$$

$$\delta_j \sim \text{N}(\zeta_0 + \sum_{i'=1}^{n^{\text{indiv vars}}} z_{j[i],i'}\delta_{i'}, \sigma_\delta^2)$$

$$\eta_l \sim \text{N}(\nu_0 + \sum_{i'=1}^{n^{\text{election vars}}} w_{l[i],i'}\nu_{i'}, \sigma_\nu^2)$$

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. (2012). Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. *Political Analysis*, *20*(04), 437–459. http://doi.org/10.1093/pan/mps023

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

Assembly, C. G. (n.d.). TABOR. *Colorado General Assembly*. Retrieved from https://public.tableau.com/views/TABOR/TABORDash?:showVizHome=no:embed=y&:display_count=no

Barr, C. D., Diez, D. M., Wang, Y., Dominici, F., & Samet, J. M. (2012). Comprehensive Smoking Bans and Acute Myocardial Infarction Among Medicare Enrollees in 387 US Counties: 1999–2008. *American Journal of Epidemiology*, *176*(7), 642–648. http://doi.org/10.1093/aje/kws267

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [Stat]*. Retrieved from http://arxiv.org/abs/1406.5823

Bergman, E., & Yates, P. A. (2011). Changing Election Methods: How Does Mandated Vote-By-Mail Affect Individual Registrants? *Election Law Journal: Rules, Politics, and Policy*, *10*(2), 115–127. http://doi.org/10.1089/elj.2010.0079

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. (2000). Voter Turnout and the National Election Studies. *Political Anal-*

ysis, *8*(4), 389–398. http://doi.org/10.1093/oxfordjournals.pan.a029823

Burden, B. C., & Kimball, D. C. (1998). A New Approach to the Study of Ticket Splitting. *The American Political Science Review*, *92*(3), 533–544. http://doi.org/10.2307/2585479

Burden, B. C., Canon, D. T., Mayer, K. R., & Moynihan, D. P. (2014). Election Laws, Mobilization, and Turnout: The Unanticipated Consequences of Election Reform. *American Journal of Political Science*, *58*(1), 95–109. http://doi.org/10.1111/ajps.12063

Bureau, U. C. (2010). US Census Bureau QuickFacts: Colorado. Retrieved from https://www.census.gov/quickfacts/co

Campbell, A. L. (2002). Self-Interest, Social Security, and the Distinctive Participation Patterns of Senior Citizens. *American Political Science Review*, *96*(3), 565–574. http://doi.org/10.1017/S0003055402000333

Chen, J. (2013). Voter Partisanship and the Effect of Distributive Spending on Political Participation. *American Journal of Political Science*, *57*(1), 200–217. http://doi.org/10.1111/j.1540-5907.2012.00613.x

Cronin, T. E., & Loevy, R. D. (2012). *Colorado Politics and Policy: Governing a Purple State*. Lincoln, UNITED STATES: UNP - Nebraska Paperback. Retrieved from http://ebookcentral.proquest.com/lib/reed/detail.action?docID=1034959

Deufel, B. J., & Kedar, O. (2010). Race And Turnout In U.S. Elections Exposing Hidden Effects. *Public Opinion Quarterly*, *74*(2), 286–318. http://doi.org/10.1093/poq/nfq017

Edelman, G., & Glastris, P. (2018). Analysis Letting people vote at home increases voter turnout. Here's proof. *Washington Post*. Retrieved from https://www.washingtonpost.com/outlook/letting-people-vote-at-home-increases-voter-turnout-heres-proof/2018/01/26/d637b9d2-017a-11e8-bb03-722769454f82_story.html

Edlin, A., Gelman, A., & Kaplan, N. (2007). Voting as a Rational Choice: Why and How People Vote To Improve the Well-Being of Others. *Rationality and Society*, *19*(3), 293–314. http://doi.org/10.1177/1043463107077384

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

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1 edition). Cambridge ; New York: Cambridge University Press.

Gerber, A. S., Huber, G. A., & Hill, S. J. (2013). Identifying the Effect of All-Mail Elections on Turnout: Staggered Reform in the Evergreen State<a

href="#fn2606">*. *Political Science Research and Methods*, *1*(1), 91–116. http://doi.org/10.1017/psrm.2013.5

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

Gronke, P., & Miller, P. (2012). Voting by Mail and Turnout in Oregon: Revisiting Southwell and Burchett. *American Politics Research*, *40*(6), 976–997. http://doi.org/10.1177/1532673X12457809

Hamm, K. (2017). How Colorado has voted in presidential elections (and how its politics have changed) since 1980. *The Denver Post*. Retrieved from https://www.denverpost.com/2017/12/22/how-colorado-votes/

Hersh, E. D. (2015). *Hacking the Electorate: How Campaigns Perceive Voters*. New York, NY: Cambridge University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: With Applications in R* (1st ed. 2013, Corr. 7th printing 2017 edition). New York: Springer.

Keele, L., & Titiunik, R. (2017). Geographic Natural Experiments with Interference: The Effect of All-Mail Voting on Turnout in Colorado.

Martin, C. (1962). *Colorado politics* (2ond ed.). Denver, Colorado: Big Mountain Press. Retrieved from http://hdl.handle.net/2027/mdp.39015024371158

Matsusaka, J. G., & Palda, F. (1999). Voter turnout: How much can we explain? *Public Choice*, *98*(3-4), 431–446. http://doi.org/10.1023/A:1018328621580

McDonald, M. P. (n.d.). What is the voting-age population (VAP) and the voting-eligible population (VEP)? *United States Elections Project*. Retrieved from http://www.electproject.org/home/voter-turnout/faq/denominator

Mettler, S., & Stonecash, J. M. (2008). Government Program Usage and Political Voice*. *Social Science Quarterly*, *89*(2), 273–293. http://doi.org/10.1111/j.1540-6237.2008.00532.x

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

Plutzer, E. (2002). Becoming a Habitual Voter: Inertia, Resources, and Growth in Young Adulthood. *The American Political Science Review*, *96*(1), 41–56. Retrieved from https://www.jstor.org/stable/3117809

Richey Sean. (2008). Voting by Mail: Turnout and Institutional Reform in Oregon*. *Social Science Quarterly*, *89*(4), 902–915. http://doi.org/10.1111/j.1540-

6237.2008.00590.x

Robert Nay. (2002). The Help America Vote Act of 2002.

Rosenstone, S. J. (2003). *Mobilization, participation, and democracy in America*. New York: Longman.

Schneider, A., & Ingram, H. (1990). Behavioral Assumptions of Policy Tools. *The Journal of Politics*, *52*(2), 510–529. http://doi.org/10.2307/2131904

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

State Legislatures, N. C. of. (n.d.). Absentee and Early Voting. *National Council of State Legislatures*. Retrieved from http://www.ncsl.org/research/elections-and-campaigns/absentee-and-early-voting.aspx#a

Stein, R. M., & Vonnahme, G. (2008). Engaging the Unengaged Voter: Vote Centers and Voter Turnout. *The Journal of Politics*, *70*(2), 487–497. http://doi.org/10.1017/S0022381608080456

Thompson, J. (2016). The first Sagebrush Rebellion: What sparked it and how it ended. Retrieved from https://www.hcn.org/articles/a-look-back-at-the-first-sagebrush-rebellion

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R* (1 edition). Boca Raton, FL: Chapman; Hall/CRC.

Wood, S., & Scheipl, F. (2017, July). Gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'. Retrieved from https://CRAN.R-project.org/package=gamm4