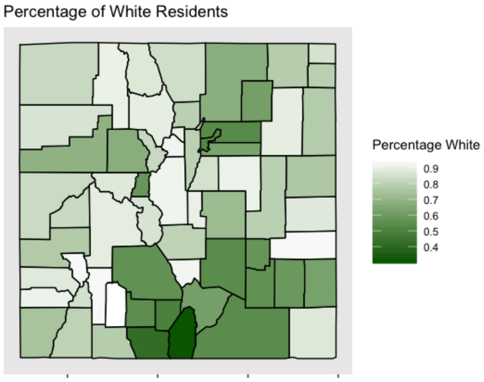

Figure 1: Map of Percentage of White Residents Per County

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.

The Centennial State and Its Voters

Demographics and Characteristics

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.

Continuing with demographic characteristics, 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 on the following map. [@census_data_2010]

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

County	Total Population	CO Population %	Largest Metro Area
Adams	441603	0.0878079	Denver-Aurora-Lakewood Metro Area
Arapahoe	572003	0.1137365	Denver-Aurora-Lakewood Metro Area
Boulder	294567	0.0585714	Boulder
Denver	600158	0.1193348	Denver
Douglas	285465	0.0567616	Denver-Aurora-Lakewood Metro Area
El Paso	622263	0.1237301	Colorado Springs
Jefferson	534543	0.1062880	Denver-Aurora-Lakewood Metro Area
Larimer	299630	0.0595781	Fort COllins
Other	1378964	0.2741917	

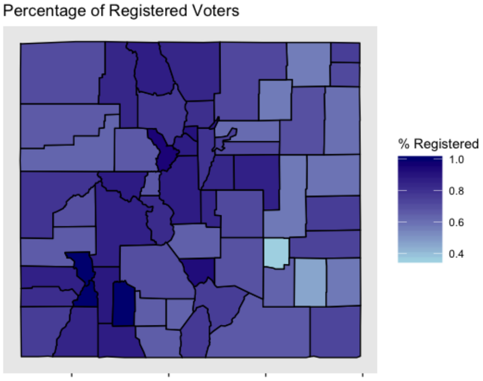

Figure 2: Map of Registration Rates Per County

County	Total Population	CO Population %	Largest Metro Area
Colorado	5029196	100.0000000	

Voting in Colorado

Each County individually administrates 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 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 2.

County	Total Registered Voters	County Voter Registration Rate	% of Statewide Registrants
Adams	270303	0.612095026528352	0.0723838
Arapahoe	410546	0.717733997898612	0.1099391
Boulder	237091	0.804879704787026	0.0634900
Denver	450616	0.750828948376927	0.1206694
Douglas	237659	0.832532884942112	0.0636421
El Paso	445708	0.716269487338955	0.1193551
Jefferson	422362	0.790136621375642	0.1131033
Larimer	250626	0.836451623669192	0.0671145
Other	1009392	—	0.2703027
Colorado	3734303	—	100.0000000

In terms of Party registration, Colorado as a whole leans democratic by a very narrow margin. This is also reflected in the state's Cook Partisan Voting Index of D +1, making it a solidly purple battleground state (Figure 3).

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

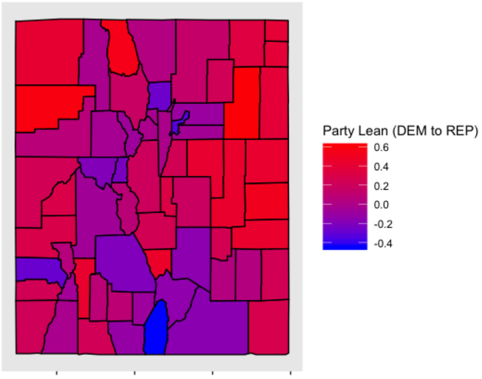

Figure 3: Map of Party Affiliation Per County

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

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

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. It gives researchers access to approximatelly 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.

Getting the Data

This thesis relies on county and individual level models to draw conclusions on voting behaviours, and how they are affected by voting method. As such, the data I need will optimally contain the following:

- **County and individual level demographic characteristics**: race, gender, urban population
- **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.

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-wode 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

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 uniquelly 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 refference 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.

Wrangling the Data

The process of "wrangling" refers to manipulating the data into a form that can then be used fo 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.

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.

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 the following graph for the eight most populous counties as described

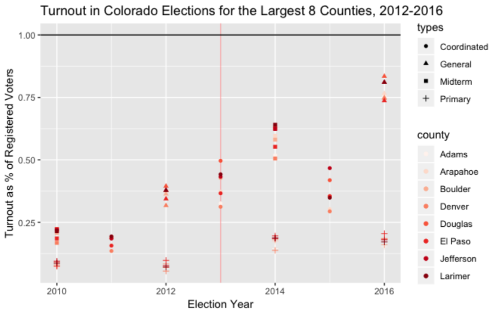

Figure 4: Turnout Plot for Eight Largest Colorado Counties, 2012-2016

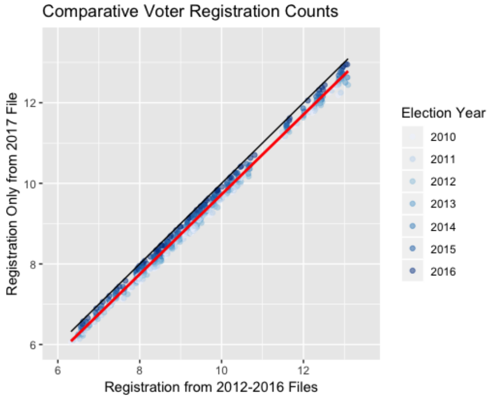

Figure 5: Comparison of registration count methods

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 this plot. 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 the following graph, 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. My selection was later vindicated, when looking at comparisons between reported rates of turnout[1] and turnout calculated through my dataset for the 2014 midterm election:

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" occured, and other similar factors.

[1] Turnout is calculated over all registered voters

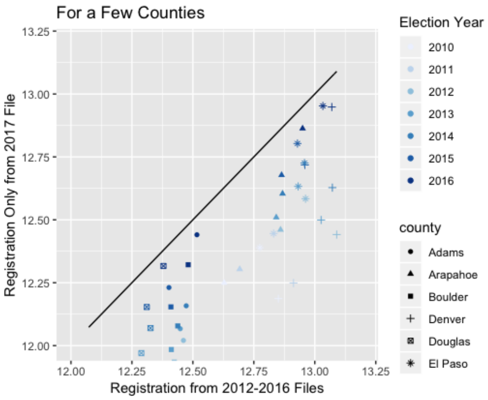

Figure 6: Comparison of registration count methods only for a few counties, 2012-2016

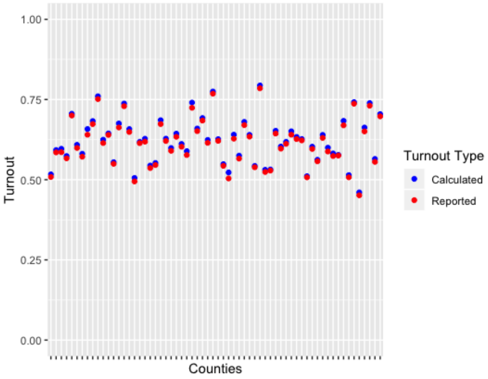

Figure 7: Comparison of reported and calculated turnout for 2014 midterms across county

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 occassionally 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 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 occured 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 likelly 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.

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 the following 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

Model Parametrization

Notation for predictors

There are four distinct types of predictors for use in these models.

County and County-per-Election Level

First, I define the following indicator variables:

• $x_c,\ for\ c \in [1, 64]$, dummy variables for each county in Colorado.

Furthermore, I have two county-level predictors:

• $x^{white\ \%}$, a vector of length 64, percentage of county population that identifies as only white.

• $x^{urban\ \%}$, a vector of length 64, percentage of county population living in an urban area.

There are two other predictors, varying by county and election. These are of particular interest, as one is the response variable for my county-level models, and the other is the variable of interest for this study. Specifically:

• $x^{mail\ vote\ \%}$, a vector of percentage of votes that was cast using mail ballots, per county and election.

• $y^{turnout\ \%}$, a vector with turnout counts per county and election. Coded with a y to identify as a response variable

Since the unit of observation for the county level models I will apply are all counties per election, I define an aggregate matrix of length equal to the number of elections times 64–the number of counties—, and width equal to 3. This matrix includes all county level predictors: $X = (x^{white\ \%}, x^{urban\ \%}, x^{mail\ vote\ \%})$. Note that this matrix includes percentage of mail ballots cast, which is the variable whose coefficient I am interested in testing.

Election Level

There are two exclusivelly election-level dicrete variables: year, and type of election. For both I define a series of indicator variables:

- $w^{election\ type}$, for each election type (Midterm, Primary , Coordinated, General).

- $w^{election\ year}$, for each election year, between 2010 and 2016.

I will also use *year* as a variable for models using smoothing splines. All election level predictors will be summarized for the purposes of modelling in the 9 by 2 matrix $W = (w^{election\ type}, w^{election\ year})$.

Individual and Individual-per-Election Level

The two aforementioned predictors–urban population and race–could be defined as aggregates of individual level observations. I also have five other distinct individual level variables:

- z^{gender}, a vector of discrete gender identifications for each voter, varying only by voter.

- z^{age}, a vector of age for each vote, varying by voter and election.

- z^{party}, a vector of party registration for each voter, varying by voter and election. Coded as Republican, Democratic, Other, or Unaffiliated.

- z^{voted}, with $z^{voted}_{i,j} = 1$ if person i voted in election j, and $z^{voted}_{i,j} = 0$ if they did not.

- $z^{mail\ ballot}$, a vector of binary values depending on whether voting method was by mail for each voter, varying by voter and election. Coded 0 if the individual did not vote.

Since the unit of observation for the individual level models I will apply are all individuals in a particular election, I define an aggregate matrix of length equal to the total number of voters, and width equal to 4. This matrix includes all individual level predictors: $Z = (z^{gender}, z^{age}, z^{party}, z^{mail\ ballot})$. The fourth variable defined in this section is the response variable in the individual level model, and as such is not included in the predictors.

County Level Models

Model 1 is a fixed-effects, bare-bones model that exclusivelly includes percentage of VBM votes, and dummy variables for year, election type, and county. Its call would look a bit like:

$$y^{turnout\ \%}_{c,l} \sim x^{mail\ vote\ \%}_{c,l}\beta_1 + \sum_{k=1}^{4} w^{election\ type}_{k,l}\beta_{k+1} + \sum_{j=1}^{7} w^{election\ year}_{j,l}\beta_{j+5} + \sum_{c=1}^{64} x_c\beta_{c+13}$$

Where k sums over the four types of election, j sums over years between 2010 and 2016, c sums over counties, and l sums over elections

Model 2 A more informed baseline, model 1 plus variables of urban and white population:

$$y^{turnout\ \%}_{c,l} \sim x^{mail\ vote\ \%}_{c,el}\beta_1 + \sum_{k=1}^{4} w^{election\ type}_{k,l}\beta_{k+1} + \sum_{j=1}^{7} w^{election\ year}_{j,l}\beta_{j+5} +$$

$$\sum_{c=1}^{64} x_c\beta_{c+13} + x^{white\ \%}_{c}\beta_{78} + x^{urban\ \%}_{c}\beta_{79}$$

This would be the “individual” level model from Gelman and Hill. I’m unsure what the “group” level for county would be. Maybe that part of the book would be more helpful for discerning effects on people’s individual p-vote?

Maybe more informative is what I did with exercise 12.2. The model tries to predict the concentration of a particular chemical based on treatment of children across time. Therefore the two levels are a visit by one individual child (here an election! so type, vbm_pct, year) and predictors for that individual child that are stable across time, like treatment type, or demographics (here race and urban pop per county).

This means I can fit a model only based on election facts, with a variable for county (models 1,3) or one that takes into account stable characteristics of the county (models 2, 4).

Model 3 A mixed-effects version of model 1, just adds mixed effects for county:

$$y_{c,l}^{turnout\ \%} \sim x_{c,l}^{mail\ vote\ \%}\beta_1 + \sum_{k=1}^{4} w_{k,l}^{election\ type}\beta_{k+1} + \sum_{j=1}^{7} w_{j,l}^{election\ year}\beta_{j+5} + \alpha_{[c],l}$$

$$\alpha_{[c],l} \sim N(0, \sigma_\alpha^2)$$

Model 4 A mixed-effects version of model 2:

$$y_{c,l}^{turnout\ \%} \sim x_{c,l}^{mail\ vote\ \%}\beta_1 + \sum_{k=1}^{4} w_{k,l}^{election\ type}\beta_{k+1} + \sum_{j=1}^{7} w_{j,l}^{election\ year}\beta_{j+5} + \alpha_{[c],l}$$

$$\alpha_{[c]l} \sim N(x^{white\ \%}\gamma_1 + x^{urban\ \%}\gamma_2, \sigma_\alpha^2),\ \ for\ c = 1, ..., 64$$

Where D is a 2 x 64 matrix of the county level predictors and γ a vector of coefficients for the county-level regression.

Model 5 During one of my discussions with Andrew, we discussed the possibility of making a model that answers the question: "Does VBM affect counties with some particular characteristic *for which I don't have data* more than others?" As such, this model would substitute county-level effects with a set of 3-4 dummy variables created through my intuitive understanding of Colorado politics and counties. For example, maybe a distinciton between central Colorado urban counties, East Colorado plains counties, and West Colorado mountain counties. The model would look a bit like:

$$y_{c,l}^{turnout\ \%} \sim x_{c,l}^{mail\ vote\ \%}\beta_1 + \sum_{k=1}^{4} w_{k,l}^{election\ type}\beta_{k+1} +$$
$$\sum_{j=1}^{7} w_{j,l}^{election\ year}\beta_{j+5} + \sum_{c=1}^{n} x_c^{county\ classification}\beta_{c+13}$$

Where $x_c^{county\ classification}$ are n dummy variables, one for each county classification group.

Model 6 As a check on the previous model, run a Principle Components Analysis on full demographic data from the 2010 census, to classify counties in the same number of groups. This model would be expected to *massively overfit*. Learning experience for all!

Note All models can work as General Additive Models with some sort of non-linear smoothing function for year. Just replace $\sum_{j=1}^{7} w_j^{election\ year}\beta_{j+5}$ with $ns(year)$.

Model 7 In order to test the hypothesis that voting by mail varies by election type, I can also construct the following model, based on model 4:

$$x_{c,l}^{mail\ vote\ \%} \sim \sum_{k=1}^{4} w_{k,l}^{election\ type}\beta_k + \sum_{j=1}^{7} w_{j,l}^{election\ year}\beta_{j+4} + \alpha_{[c]}$$

$$\alpha_{[c],l} \sim N(x^{white\ \%}\gamma_1 + x^{urban\ \%}\gamma_2, \sigma_\alpha^2),\ for\ c = 1, ..., 64$$

This would predict whether there are specific county or election characteristics that increase the amount of mail ballots individuals cast.

Individual Level Models

This section follows directly from the intro to Gelman & Hill's 11th chapter.

Model 8 As a baseline for all further analysis, a logistic regression that treats each vote in a single election as uniform across counties, as such not including any group-level predictors.

$$P(z_{i,l}^{voted} = 1) = logit^{-1}(Z_{i,l}\delta + W_l\beta)$$

Where matrices Z, Y are as described above, i is an indice for each voter, and l for each election. δ, β are vectors of coefficients to be estimated.

Model 9 Add group level mixed effects and predictors.

$$P(z_{i,l}^{voted} = 1) = logit^{-1}(Z_{i,l}\alpha + W_l\beta + \alpha_{[c],l})$$

$$\alpha_{[c],l} \sim N(X_c\gamma, \sigma_\alpha^2),\ for\ c = 1, ..., 64$$

Where X_c as defined above, and γ a vector of coefficients.

Model 10 Include extra model with EM algorithm applied to 2015 data maybe?