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.

Hypotheses

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?*

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 competing voter choice theories; 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 incremental 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*

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.

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 important, 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 quite important, 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.

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.

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:

$$f(Y) = X\beta,$$

where Y is a vector of response variable values, X is a matrix of predictors, and B is a vector 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 [1]. 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:

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

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 vecotr will not be particularly helpful, e^B can be used as a vector of multiplicative, one-unit shifts in the value of the probability that $Y = 1$. 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[2]. [@james_introduction_2017]

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

[1] Or, in this case, the range of the parameter defining the distribution of the response, which is p for the Bernoulli distribution

[2] This can be simplified even more, if one approximates the process of exponentiating with just dividing the coefficient by 4. Crude, yet effective for a quick scan of the results

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 fewer 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.

One solution is the 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_introduction_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 [@barr_comprehensive_2012]. In terms of this thesis, this will help when responding to Q2 as it was framed earlier in this chapter.

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 are estimated 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_data_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:

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

Multilevel models can be fit using the `lme4` *R* package that uses restricted maximum likelihood calculations for estimating coefficients [@bates_fitting_2014]. Multilevel modelling also works perfectly well with general additive models! In *R* this can be accomplished with the `gamm4` package [@wood_gamm4:_2017].

Model Accurracy and Quality of Fit

Mean Squared Error (MSE)

For all generalized linear regression models (including GAMs, mixed and fixed effects models) I use Mean Squared Error to asses the accuracy of the fit. Assuming a dataset $\{(y_0, x_0^1, x_0^2, ..., x_0^p), ..., (y_n, x_n^1, x_n^2, ..., x_n^p)\}$ of n observations and p predictors, with X_i a vector of the predictors for the i-th observation, and $f : R^n \to R$

the true multivariate function connecting the predictors and response, mean squared error is calculated as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(X_i))^2$$

MSE can be calculated either using the same dataset used in estimating the model coefficients, or on a new dataset. In the later case it is called predictive or test MSE. Despite prediction not being the purpose of the models presented in this thesis, I use test MSE because of the independence such a calculation method brings from the data used for the fit, compensating in a way for overfitting[@james_introduction_2017]. To calculate test MSE I use five-fold cross-validation, which will be analyzed shortly.

Area Under the Curve (AUC)

Logistic regression models estimate the probability of a binary variable taking a TRUE value. The predictive output of such a model will be a series of probabilities. These probabilities can then be used to approximate a dataset of positive and negative values for the response variable (in my case, voting). Based on the true values of the response, one can calculate the counts of true positive, true negative, false positive, and false negative predictions. To make this calculation, a probability threshold is set over which the prediction for the response is positive. Positive predictive values of the response are assigned based on the following statement:

$$\text{P}(y_i = 1|X_i) > p$$

where y_i, X_i can be assumed to be the same as in the previous section, and p is the threshold. A common and intuitive threshold is 0.5, but any number in $(0, 1)$ can be used. After getting counts for true/false negative/positive values, one can then calculate *specificity* and *sensitivity* for the model. These are:

$$\text{specificity} = \frac{\text{True Positive}}{\text{False Negative + True Positive}}$$

$$\text{sensitivity} = \frac{\text{True Negative}}{\text{False Positive + True Negative}}$$

Using these three metrics–sensitivity, specificity, and probability threshold–it's possible to create an ROC curve, which is one of the most widely used diagnostic plots for classification models[3]. The ROC curve has $1 -$ specificity on the x-axis, sensitivity on the y-axis, and each point describes a pair of x-y values for each value of the probability threshold. Using this plot, it's possible to measure the *area under the (ROC) curve*, or AUC, which serves as a goodness-of-fit measure for classification models. The AUC is a number in the $[0, 1]$ range and should be maximized; a .5 AUC is representative of an ROC curve on the $y = x$ line, which is a coin-toss no-information classifier. Plot 2.1 is an example of an ROC curve.

Similarly to MSE, there is value in calculating AUC from a test dataset, rather than the dataset used to train the model. Therefore I also use 5-fold cross-validation for AUC as well[4].

k-Folds Cross Validation

The goal of statistical modeling is to approximate the true function that links predictors to response. While the final model's coefficients should be estimated using as much data as possible, when assesing how good a fit that model is there can be better uses of the power that large amounts of data give us. k-Folds cross validation allows for better approximations of goodness-of-of-fit metrics, by partitioning the data into training datasets and test datasets. The fundamental idea is that the data is split into k different subsets, which are then sequentially used to fit the model and calculate the value of some metric. The algorithm goes as follows:

[3]The ROC curve takes its name from a term in communications science, the *receiver operating characteristics curve*. The name is historic, and not relevant to its statistical application.

[4]This also compensates for models not converging, as some of mine do.

1. Partition data into k folds
2. Fit model on all but the i-th fold
3. Calculate goodness-of-fit metric on the i-th fold
4. Repeat 2 and 3 for i$\in [0, k]$
5. Calculate the average of all obtained goodness-of0fit measurements