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. There are two types of models that will be included here: county-level, and individual-level. They will be treated as seperate sections.

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.

County Level Models

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 assignement of voting tendencies across counties; each county independent of election has a unique range of turnout results. With the addition of some county-level characteristics, it is possible to build a naive, baseline model of turnout as follows:

$$y_{k,l}^{turnout} = \beta_0 + (\sum_{k=1}^{64} \beta_k x_k^{county}) + \beta_{65} x_k^{\%white} + \beta_{66} x_k^{\%urban},$$

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_data_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 modelled 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 [@wood_generalized_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.[1] A summary of these four models is provided in the following table:

INSERT TABLE

Individual Level Models

[1] 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.