Model Specs—3 Weeks Out

Theodore Dounias

10/2/2018

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:

- $y^{mail\ vote\ \%}$, 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.

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:

- $y_i^{election\ type},\ for\ i \in [1, 4]$, for each election type (Midterm, Primary , Coordinated, General).
- $y_i^{election\ year},\ for\ i \in [1, 7]$, 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 $Y = (y_i^{election\ type}, y_i^{election\ year})$.

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 four 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 voter, 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.

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 product of elections and voters, a.nd width equal to 3. This matrix includes all individual level predictors: $Z = (z^{gender}, z^{age}, z^{party})$. 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:

$$y^{turnout\ \%} \sim y^{mail\ vote\ \%}\beta_1 + \sum_{i=1}^{4} y_i^{election\ type}\beta_{i+1} + \sum_{j=1}^{7} y_j^{election\ year}\beta_{j+5} + \sum_{c=1}^{64} x_c\beta_{c+13}$$

Where C, T, Y are vectors of dummy variables and c, tp, y are indices of county, election type, and year. mail% is the percentage of the vote conducted through VBM. This model serves as a baseline for comparing others.

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

$$y^{turnout\ \%} \sim y^{mail\ vote\ \%}\beta_1 + \sum_{i=1}^{4} y_i^{election\ type}\beta_{i+1} + \sum_{j=1}^{7} y_j^{election\ year}\beta_{j+5} + \sum_{c=1}^{64} x_c\beta_{c+13} + x^{white\ \%}\beta_{78} + x^{urban\ \%}\beta_{79}$$

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:

$$y^{turnout\ \%} \sim y^{mail\ vote\ \%}\beta_1 + \sum_{i=1}^{4} y_i^{election\ type}\beta_{i+1} + \sum_{j=1}^{7} y_j^{election\ year}\beta_{j+5} + \alpha_{[c]}$$

$$\alpha_c \sim N(0, \sigma_\alpha^2)$$

Model 4 A mixed-effects version of model 2:

$$y^{turnout\ \%} \sim y^{mail\ vote\ \%}\beta_1 + \sum_{i=1}^{4} y_i^{election\ type}\beta_{i+1} + \sum_{j=1}^{7} y_j^{election\ year}\beta_{j+5} + \alpha_{[c]}$$

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

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:

$$y^{turnout\ \%} \sim y^{mail\ vote\ \%}\beta_1 + \sum_{i=1}^{4} y_i^{election\ type}\beta_{i+1} + \sum_{j=1}^{7} y_j^{election\ year}\beta_{j+5} + \sum_{c=1}^{n} x_c^{county\ classification}\beta_{c+13}$$

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 $\sum_{j=1}^{7} y_j^{election\ year}\beta_{j+5}$ with $ns(year)$.