Predicting Turnout from State Voter Files: A Proposed Method and an Evaluation of Competing Specifications

Corwin D. Smidt
Assistant Professor
Department of Political Science
Michigan State University
smidtc@msu.edu
August 31, 2015

Abstract

Predictions of voter turnout from state voter files are useful to scholars and practitioners who need an accurate sample of the electorate. I explore how scholars with even limited resources can create accurate turnout predictions by utilizing a subsampling bootstrap imputation and estimation method that puts the estimation task on a memory and computational scale suitable for desktop computers. I compare the performance of different model specifications in making out-of-sample predictions using data from the state of Michigan's Qualified Voter File (QVF) before and after the 2014 elections. Predictions based on empirical Bayes estimates of unobserved heterogeneity at the individual level consistently perform better than predictions simply based on observable behavior and demonstrate how subsampling provides a practical way for scholars to estimate complex and accurate predictive models of turnout for either local survey sampling or public use.

Prepared for presentation at the Annual Meeting of the American Political Science Association; San Francisco, CA; September 5, 2015
There is growing demand for predictions of voter turnout from state voter lists. Election surveys struggle to reach respondents and identify which of them represent the electorate. Likely voter filters perform erratically prior to election day and often fail to represent the electorate (Erikson, Panagopoulos, and Wlezien 2004). And respondents often falsely report whether they voted out of social desirability bias after the election. A promising cost-effective alternative is to probabilistically sample from state voter files based on an estimate of each registered voter’s probability of voting (Barber et al. 2014).

Indeed, beyond their academic value, micro-level forecasts of voter turnout also represent an easy way that political science can provide a public good. Candidates for local office either waste time and money targeting unlikely voters or spend a lot of money on campaign consultants to tell them whom to target. Providing accurate and accessible estimates can potentially make it easier and cheaper for local candidates to run for office.

But political science has been slow to develop and evaluate micro-level turnout predictions, at least publicly. One problem is access to data. Many states provide voter history files for free or for a very small processing fee, but most data sets are difficult to manage and require specialized knowledge of the state. After cleaning up these data, further problems emerge in the form of estimation choices and constraints with significant big data hurdles. Many states have over millions of registered voters, each with repeated observations that can extend across multiple decades. Depending on a scholar’s computing resources, this limits the complexity of the model available for estimation. Moreover, many voter histories are only specific to a current address, which potentially creates a large element of missing data and the need for the imputation of histories using individual and address-based information. But the standard practice of creating 5 or 10 imputed values is impractical when voter file observations are in the millions.

Instead of facing these challenges, it is common to see scholars rely on practitioners by acquiring voter histories and turnout predictions from campaign organizations or private political consulting firms (e.g., Gerber, Green, and Larimer 2008). Most prominently, the CCES merged its survey sample with private voter history data from the for-profit Catalist political consulting firm (Ansolabehere and Hersh 2012). This is unfortunate and likely prohibitive for future research. Acquiring private information can restrict academic projects in scope and accuracy because of costs. It also lacks scientific transparency and limits scholarly evaluation of a prediction’s accuracy and components.

In the following, I explore how scholars with even limited resources can create accurate turnout predictions. First, I explore the performance of a subsample im-
putation and estimation method. In exchange for a small element of sampling error, the smaller subsamples do not limit the content of predictors or the complexity of the model specification in the face of big or missing data. This puts the estimation task on a memory and computational scale that is manageable and quicker for most high performance multi-core desktop computers.

After validating the accuracy and negligible costs of this approach, I compare the performance of different model specifications in making out-of-sample predictions as model form and complexity changes using data from the state of Michigan’s Qualified Voter File (QVF) before and after the 2014 elections. I compare out-of-sample performance in terms of its micro-level classification abilities and its bias and efficiency at the aggregate level. I find that predictions based on empirical Bayes estimates of unobserved heterogeneity at the individual level consistently perform better than predictions from simply contextual or lagged-based predictors. These results demonstrate that subsampling estimation provides a practical way for scholars to estimate complex and accurate predictive models of turnout for either local survey sampling or public use.

**When Theory Meets Data (and Computation Capacity)**

Theories of voter participation are legion and differ in their theoretical orientation. Most acknowledge that turning out to vote is an act of varying expressive and instrumental utility that depends on a person’s resources and social learning acquired through socio-economic status, strength of partisanship, geographic mobility, and civic skills (e.g., Achen 2006; Gerber and Green 2000; Leighley and Nagler 2013; Verba, Schlozman, and Brady 1995). These studies commonly point to a core set of individual-level predictors: age, income, education, residential stability, awareness, and political orientations, like strength of partisanship or efficacy. Along similar lines, recent studies convincingly show that the prior act of voting adds to this learning experience and is habit-forming (Gerber, Green, and Shachar 2003; Plutzer 2002), where the act of voting in the prior election has independent effects on future behavior.

Only some of these theories are of use when trying to predict turnout from state voter files, which typically only record a voter’s name, age, address, prior history, and a proxy for residential stability in the form of the registration date.¹ Professional firms and organizations can buy additional information from marketing

---

¹. As parties have opened their primaries, state records of party registrations have weakened to the point they are of unclear value compared to records of turning out to vote in presidential primaries.
databases to match with voter records (Hersh 2015; Issenberg 2012), but this is not practical for most academic budgets. Indeed, the more practical approaches for academics bring with them many additional problems.

One potential approach is to include additional variables to proxy for the unmeasured attributes. Although straightforward, there are some clear limitations to this approach. Block-level census data are only available through the decennial census or the 5-year American Community Survey and do not precisely match political jurisdictions. Moreover, they are tabulated among all residents within a block, not all registered voters. Lagged turnout measures at a ward or precinct-level are also of varying use depending on when redistricting occurred. But the greatest concern is that these predictors do help scholars or candidates who need to identify who votes within a specific locality, since these geographic-based attributes cannot identify which voters are more or less likely to vote within a census block or within a voting precinct.

Including lagged indicators of past turnout allows one to sort out the likely voters at the individual-level since prior history is a powerful predictor and indicator of one's likelihood of voting. But simple estimates of prior turnout's relationship with future turnout are limited because records are only commonly available for a person's current place of residence. This biases estimates both within elections, since voting histories are missing from individuals who recently moved and tend to be of lower income or younger. And this also creates small sample bias in estimating the association across elections, since the majority of observations will only record associations between the few recent election cycles. Multiple imputation of these missing variables certainly can address the former type of bias (e.g., King et al. 2001; Schafer 1997). The practice is computationally costly, however, since it requires of the estimation of an imputation model and the estimation of a prediction model across multiple draws of the missing values.

Alternatively, there is the latent classification approach: treat a voter's history as a set of indicators of their latent tendency to turn out to vote. Although possible through an IRT or latent class model, multi-level modeling frameworks are a good entry point since they offer a little more accessibility in accounting for missing data and voting's potentially habit-forming behavior (Skrondal and Rabe-Hesketh 2004). There are also limitations and costs to this approach, but it is unclear how they compare to the alternative. Most notably, there are heavier computational burdens in estimating these panel/multi-level models within very large datasets. Quadrature-based approximations of the likelihood integrals take a lot time for
large data sets, and this time is only slightly reduced by the number of cores. Once having an estimate, there is the need to accurately calculate the empirical Bayes estimate of each voter’s unobserved, time-constant tendency to turn out to vote. Multi-level estimation does not require the imputation of missing histories, but it would provide no added-value for those cases since the estimate of the unobserved tendency would be zero.

A Bootstrapping Approach to Big Data Imputation and Prediction

Since predictive models should be chosen based on their predictive accuracy and not their computational demands, we need ways to estimate these alternative frameworks under common computational constraints. Since sampling’s original purpose was to reduce computational burdens and research costs, it seems natural to consider whether a subsampling-based bootstrap estimator offers a practical way to approximate big data inferences. Instead of RAM limits choking estimation across the full set of observations, the intuition is to take an average estimate across smaller subsamples where one can choose to carry out estimation on a scale that allows for simultaneous estimation across separate cores that is appropriate for available RAM.

Bootstrap estimators are a surprisingly effective tool for estimating confidence intervals or predictions of complex population characteristics (Efron and Tibshirani 1993; Horowitz 2001). And subsampling bootstrap estimators can be adjusted to approximate a full-scale resampling bootstrap and offer a way to reduce computational burdens (Politis, Romano, and Wolf 1999). Indeed, there are many cases where bootstrap estimate for a population size \( n \) based on smaller subsamples \( (b < n) \) work despite the failure of the \( n \) bootstrap estimator (Bickel, Götze, and Zwet 1997; Politis, Romano, and Wolf 1999).

These conditions are of little relevance in this case since the goal is to generate an average of likelihood-based parameter estimates, which fit within the assumptions for the full-scale bootstrap. Likewise, studies of subsampling focus on its ability to make bias or asymptotic adjustments to finite-sample bootstrap estimates or

2. In its most recent report, Stata reports no speed improvement between its single-core and multi-core estimation of panel mixed effects models \texttt{xtmixed} and \texttt{xtmelogit} (see www.stata.com/statamp/statamp.pdf).

3. For instance, see Kleiner et al. (2012) for discussion of their Bag of Little Bootstraps (BLB) estimator which seems akin to the ideas of nested or iterated bootstrapping to adjustments for non-asymptotically pivotal statistics (Cameron and Trivedi 2005, 374).
evaluating whether a full-sample bootstrap estimate is valid (Cameron and Trivedi 2005, 373). While potentially helpful, they are not of immediate relevance when applying it to millions of observations and numerous cities and precincts. There is no concern that likelihood's asymptotic assumptions hold within subsample sizes under consideration here. Instead, I am simply utilizing subsampling as a way to make big data prediction more manageable by reducing the computationally intensive imputation process in estimating prior voting history or multi-level effects.

In regards to imputation, since there is no immediate need to estimate standard errors, there need only be one imputation draw within each subsample estimate to generate coefficient estimates that correct for missing data bias. Let us take the example of using a lagged turnout indicator to predict future turnout. Define individual $i$'s act of turning out to vote at time $t$ by the dichotomous $y_{it}$, where we need to estimate $\Pr[y_{it+1} = 1 | y_{it}, x]$ via a logit model and $y_{it}$ is not completely observed. The proposed procedure is then as follows:

1. Take a random sample of size $b$ from the population of size $N$, with replacement, where $b < N$.
2. Assuming missing at random, estimate $\Pr[y_{i,t-1} = 1 | x] = \Lambda(x_i' \hat{\gamma})$ among non-missing observations (where $\Lambda(\cdot)$ is the logistic cumulative distribution function); save estimates $\hat{\gamma}_s$.
3. For missing observations, draw $y_{i,t-1}$ from $\text{Bern}(\Lambda(x_i' \hat{\gamma}_s))$.
4. Estimate $\Pr[y_{it} = 1 | y_{i,t-1}, x] = \Lambda(x_i' \hat{\beta} + \rho y_{i,t-1})$; save estimates $\hat{\beta}_s$ and $\hat{\rho}_s$.
5. Repeat steps 1-4, $S$ times.
6. Among all observations with $y_{it}$ observed, estimate $\Pr[y_{i,t+1} = 1 | y_{it}, x]$ by $\Lambda(x_i' \hat{\beta} + \hat{\rho} y_{it}^m)$, where $\hat{\rho} = \sum_{s=1}^{S} \hat{\rho}_s / S$ and $\hat{\beta}^k = \sum_{s=1}^{S} \hat{\beta}_s^k / S$ for each parameter $k$ in $\beta$.
7. Among all observations with $y_{it}$ unobserved:
   (a) draw $y_{it}^m$ from $\text{Bern}(\Lambda(x_i' \hat{\gamma}))$, $\hat{\gamma}^k = \sum_{s=1}^{S} \hat{\gamma}_s^k / S$ for each parameter $k$ in $\gamma$.
   (b) estimate $\Pr[y_{i,t+1} = 1 | y_{it}^m, x]$ by $\Lambda(x_i' \hat{\beta} + \hat{\rho} y_{it}^m)$, save each estimate $\hat{Pr}[y_{i,t+1}^m]$.
   (c) After $M$ draws, estimate $\Pr[y_{i,t+1} = 1 | y_{it}, x]$ by $\sum_{m=1}^{M} \hat{Pr}[y_{i,t+1}^m] / M$.

4. Schafer (1997) suggests single imputation is sufficient for even suitable standard error estimates when less than 5% of the observations are missing.
In regards to multi-level modeling, the subsampling simply needs to be stratified within each nesting level for unbiased estimates of the variance components among geographic regions or political jurisdictions. If model estimation among a random subsample of say 20 individuals within each jurisdiction remains too computationally burdensome, it is possible to follow a less efficient, multi-stage cluster sampling design and draw a random subsample of jurisdictions and then a random subsample of individuals within each jurisdiction.

**Simulation Evidence**

Past studies document the advantages of bootstrap and subsampling estimates in terms of the bias-correction, efficiency, and power of their in-sample coefficient estimates and tests (Efron and Tibshirani 1993; Politis, Romano, and Wolf 1999). These would suggest suitable performance in imputation and out-of-sample predictions, but it is instructive to conduct a simulation that measures its relative performance in concrete ways.

For each simulation, I created a data set that recorded 5 election cycles of voting history for 1 million observations. The observed voting histories are themselves generated from a latent probability that is a function of 6 observed covariates (turnout in previous cycle, age, age-squared, off-year elections, and years since registered). After estimating coefficient estimates and generating out-of-sample predictions of turnout in the 6th cycle for the full set of observations (where previous turnout is imputed for the subset of missing observations), each simulation generated a set of bootstrap estimates from multiple resamples of the full data. Each simulation generates a set of bootstrap predictions for estimators that vary by the size of the bootstrap subsample (10 conditions ranging from 1,000 - 50,000) and by the number of bootstrap draws (2 conditions: 10 and 20), for a total of 20 different bootstrap predictors. $M$ is set constant at 5 imputation draws.

Figure 1 evaluates the relative performance of the observed and bootstrap predictions relative to the latent probability of voting in the next election cycle by calculating the average root mean squared-error (RMSE) in the predicted probability and average maximum absolute error in predicted probability. As bootstrap sample sizes grow, the relative performance of the bootstrap predictions change from about 10 to 2 times as worse than the predictions from fully observed data. But the size of the prediction error seems relatively inconsequential on an absolute scale. Regardless of the number of bootstrap draws, if the subsample size is 10,000 or larger, the average maximum error is less than 1 percentage point and the RMSE in the predicted probability is less than 0.3 percentage points. These patterns suggest suitably precise estimates are available with minimal requirements in compu-
Average maximum absolute error and root mean squared error (RMSE) in the out-of-sample predicted probability across 200 simulations. Bootstrap 10 and Bootstrap 20 indicate prediction estimates respectively based on 10 and 20 subsample bootstrap draws.

More importantly, across either metric of prediction error, the performance is equal across effective number of observations. Doubling an estimator’s bootstrap samples reduces prediction error equal to doubling its subsample size. This makes the bootstrap approach much quicker than the full sample estimate if separate subsample estimates can be conducted across multiple cores at the same time.

These simulations demonstrate that subsampling can be an effective way to generate accurate predictions with less computational burden, but it is unclear if this translates into suitable performance as model specifications change, especially...
when applied to real data; a task I now turn to.

**Application: Predicting 2014 Turnout from the Michigan State Qualified Voter File**

I evaluate and compare the performance of different predictive models of voter turnout within the 2014 and 2015 Michigan QVF. I focus on Michigan because of my familiarity with the data and the state, but it is an instructive state to analyze since it has a diverse set of regions in terms of race, income, and residential mobility. I assess performance in predicting turnout for the relatively uncompetitive August 2014 primary and the November general, which featured a competitive governor’s race and a somewhat lackluster campaign for an open U.S. Senate seat. I generate primary and general election turnout probabilities for the 7.4 million registered voters listed within the June 2014 Michigan QVF and compare those predictions to the official turnout records at the county level and within the January 2015 Michigan QVF.

Voters who have changed their residence between the dates of the June 2014 QVF and January 2015 QVF will not have a record of turnout since the state changes a voter’s identification number each time he or she registers at a new address. This makes it impossible to connect a voter’s information to their prior identification number which the state would have used to record turnout in prior elections. 13.3% of voters in the 2014 QVF have no record of behavior for 2012 and only about half of Michigan voters have traceable records through 2004. Limiting my analysis to these 5 cycles still constitutes 48.9 million total observations across primary and general elections in even years.

In the following, I compare the performance of different specifications by summarizing estimates across 20 bootstrap subsamples of 100,000 voters from the QVF, which results in approximately 670,000 turnout observations for each bootstrap subsample

**The “Just Add Variables” Approach**

As discussed, one possible approach to prediction is to regress on the variables provided within the QVF and other contextual attributes that can be matched to the voter’s name or address. I start with a *basic* model that only predicts based on

---

5. I choose to ignore the role of off-year elections. Michigan has the most decentralized election administration in the U.S., with 1,599 separate election administrators. The QVF records when a voter voted in an off-year or off-cycle election, but the state does not keep of whether a voter who did not turn out had the opportunity to vote on that date.
information within the QVF, where:

\[
\Pr[y_{ijt} = 1] = \Lambda(x'_{ijt}\beta + \phi Comp_t)
\]

for \(I\) individuals across \(J\) jurisdictions and \(T\) elections. \(x_{ijt}\) is a set of fully observable predictors that includes age (natural cubic spline, plus interactions by election type), gender, log of how many months registered at that address, and a set of election dummies (August vs. General; Gov. vs. Pres. Election). \(Comp_t\) proxies for the effects that statewide competitiveness might have on behavior by using a measure the observed turnout across the state. Since 2014 turnout is already observed this is cheating within this context, but the idea is that this would be set by scholar prior to the election in recognition of the amount of competition at the top of the ballot.

I then add a precinct lag measure of turnout to evaluate how much contextual attributes of a voter's address can improve prediction.

\[
\Pr[y_{ijt} = 1] = \Lambda(x'_{ijt}\beta + \phi Comp_t + \sum_l \gamma_l z_{jt-l})
\]

where \(z_{jt-l}\) represents the turnout rate in a prior election of voters currently in same precinct who also lived in that precinct in that prior election (centered to state-wide turnout in that election). I include both two and four-year lags that are specific to each type of election (primary or general). In practice this might be supplemented with other measures taken from census data, but a precinct level measure provides sufficient leverage in identifying areas that would likely have higher or lower levels of turnout.

The problem with a precinct lag model is that it does not provide sufficient differentiation among voters within each precinct. The third model adds to the previous two by including a set of vote lag indicators \((y_{ijt-l})\) for each voter.

\[
\Pr[y_{ijt} = 1] = \Lambda(x'_{ijt}\beta + \phi Comp_t + \sum_l \gamma_l z_{jt-l} + \sum_l \rho_l y_{ijt-l})
\]

Like the precinct lags, I include both two and four-year lags that are specific to each type of election (primary or general). The addition of the four-year lag pushes the missing observation quotient to around 30%, but it is necessary since turnout in presidential and midterm elections are not strongly associated with each other.

Table 1 compares each model's performance in distinguishing between voters and counties that were and were not likely to turnout out to vote. The top panel in Table 1 compares each model's performance in making predictions at the individual level by comparing the area under the receiver operating characteristic (aROC)
Table 1. Comparing Sorting Capabilities of Model Predictions

<table>
<thead>
<tr>
<th>Model</th>
<th>Election</th>
<th>August</th>
<th>November</th>
</tr>
</thead>
<tbody>
<tr>
<td>aROC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td>.738</td>
<td>.682</td>
</tr>
<tr>
<td>Precinct Lag</td>
<td></td>
<td>.742</td>
<td>.690</td>
</tr>
<tr>
<td>Vote Lag</td>
<td></td>
<td>.847</td>
<td>.811</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>County-level correlation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td></td>
<td>.407</td>
<td>.267</td>
</tr>
<tr>
<td>Precinct Lag</td>
<td></td>
<td>.613</td>
<td>.882</td>
</tr>
<tr>
<td>Vote Lag</td>
<td></td>
<td>.639</td>
<td>.901</td>
</tr>
</tbody>
</table>

County-level correlation calculated between estimated and actual turnout rate among registered voters across the 83 Michigan counties.

curve for turnout within the August primary and November general election. All models do slightly better in classifying likely voters for the August primary rather than the November general. The basic logit does only a fair job of distinguishing which voters are likely to turnout. Not surprisingly, the precinct lag model does not offer enough differentiation at the individual-level to provide substantial improvement. In comparison, and despite its high imputation demands, including both a two-year and four-year lag of voter turnout offers substantially better performance in distinguishing likely and unlikely voters at the individual level. It pushes the aROC to a respectable .85 and .81 for the respective August and November elections.

When looking at each model’s ability to predict the low and high turnout counties, we find the basic model not doing well at all. Its predictions of county turnout rates only correlate at a rate of .27 with turnout observed during the November general. The value of adding precinct-level regressors is much more apparent in making aggregate predictions, however. The simple addition of precinct lags boosts county-level prediction correlations to .613 in August and .882 in November. In comparison, lagged indicators of a person’s prior turnout provides only slight improvements in classifying the high and low turnout counties for 2014, pushing the
Table 2. Comparing County-level Accuracy of Model Predictions

<table>
<thead>
<tr>
<th>Model</th>
<th>Election</th>
<th>August</th>
<th>November</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>Bias</td>
<td>4.96</td>
<td>4.73</td>
</tr>
<tr>
<td>Precinct Lag</td>
<td>Bias</td>
<td>5.56</td>
<td>5.42</td>
</tr>
<tr>
<td>Vote Lag</td>
<td>Bias</td>
<td>-0.19</td>
<td>-6.50</td>
</tr>
<tr>
<td>Basic</td>
<td>County-level RMSE</td>
<td>6.47</td>
<td>6.46</td>
</tr>
<tr>
<td>Precinct Lag</td>
<td>County-level RMSE</td>
<td>6.85</td>
<td>5.84</td>
</tr>
<tr>
<td>Vote Lag</td>
<td>County-level RMSE</td>
<td>3.61</td>
<td>6.79</td>
</tr>
</tbody>
</table>

Bias and RMSE are expressed in terms percentage point difference in turnout rate among registered voters across the 83 Michigan counties.

It is impressive to find correlations in Table 1 that approach .9 between predicted and observed county turnout rates. But these correlations do not demonstrate the substantial degree of estimation bias. The top panel in Figure 2 shows which counties had turnout rates higher and lower than expected for each model. Since the less populous counties tend to have errors in turnout rates that are of greater magnitude, the bottom panel compares county errors in terms of difference in total number of voters. The performance in November is consistently poor. Turnout across all counties was much lower than expected by either the basic or precinct lag model. Although the vote lag model's predictions are relatively unbiased in August, they oddly provide a severe underestimate of turnout in November. What makes this poor performance even more striking is that they already include the observed statewide turnout rate in 2014 to account for campaign competitiveness in each year.

Table 2 summarizes these error rates by presenting the average bias and root mean squared error (RMSE) for each county turnout rate prediction. The average error in November ranges from 5.4 for the precinct lag to -6.5 for the vote lag; that is a big difference. And, according to the RMSE, the typical prediction error is near or over 6 percentage points across most models. There are many possible
Figure 2. County-level Errors in Turnout Rate (top) and Turnout Count (bottom) Estimates
explanations for these patterns, but a general explanation is that they may not be that effective at predicting turnout rates when they are based on observations across a few recent cycles. Indeed, perhaps the most striking pattern is the sensitivity of the turnout estimates to what variables are included. In this case, the inclusion of a couple regressors at the individual level substantially flipped prediction errors for the next cycle.

The Classifying Voter (Empirical Bayes) Approach

Instead of using previous votes as predictors of future voting behavior, an alternative approach is to use voter histories as indicators of the type of voter each person is. Based on the model estimates of the distribution or commonality of each type of voter, classification estimates can be generated for each voter using an empirical Bayes estimate (Efron and Morris 1973). For ease of comparison, I start with a basic random effects panel logit, since it can be easily integrated within the previous basic model.

$$\Pr[y_{ijt} = 1] = \Lambda(x'_{ijt}\beta + \phi\text{Comp}_i + u_i)$$

where $u_i$ represents the individual-specific time constant error term that is independent of $x_{ijt}$ and distributed normally with mean zero and estimated variance $\sigma_u^2$. The empirical Bayes estimate compares each individual’s actual record of turnout with turnout expected under the observed $x'_{ijt}\beta$ component to estimate $u_i$ via Bayes’ theorem.

The inclusion of the lagged precinct parameters are a simple addition to this model, but the inclusion of the vote lag measure for each person makes that model slightly more complicated in form. Namely, to address the initial conditions problem within dynamic panel models I rely on the Wooldridge (2005) simple solution, which conditions the dynamic estimate based on the first observation of $y$ ($y_{ij1}$).

$$\Pr[y_{ijt} = 1] = \Lambda(x'_{ijt}\beta + \phi\text{Comp}_i + \sum_l \gamma_lz_{jt-l} + \sum_l \rho_ly_{ijt-l} + \alpha y_{ij1} + u_i)$$

Even if voting is largely habit forming and can change a person's long-standing predilection to participate (i.e., Gerber, Green, and Shachar 2003), the addition of these lagged measures may be too burdensome for efficient turnout predictions.

6. One of the more likely explanations for the vote lag model’s poor performance is that the observed association between turnout in 2002, 2006, and 2010 midterms must have been much weaker than the association between turnout across 2010 and 2014. Those three prior midterms alternated between Republican, Democratic, and Republican landslide elections, which may have made for a rotating base of turnout. But the 2010 and 2014 midterms were much more similar in their Republican appeal, perhaps driving a similar group of voters to each poll.
Table 3. Comparing Sorting Capabilities of Panel Model Predictions

<table>
<thead>
<tr>
<th></th>
<th>Election Model</th>
<th>August</th>
<th>November</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>aROC</strong></td>
<td>Basic</td>
<td>.863</td>
<td>.817</td>
</tr>
<tr>
<td></td>
<td>Precinct Lag</td>
<td>.864</td>
<td>.820</td>
</tr>
<tr>
<td></td>
<td>Vote Lag</td>
<td>.859</td>
<td>.818</td>
</tr>
<tr>
<td><strong>County-level correlation</strong></td>
<td>Basic</td>
<td>.678</td>
<td>.810</td>
</tr>
<tr>
<td></td>
<td>Precinct Lag</td>
<td>.677</td>
<td>.897</td>
</tr>
<tr>
<td></td>
<td>Vote Lag</td>
<td>.649</td>
<td>.744</td>
</tr>
</tbody>
</table>

County-level correlation calculated between estimated and actual turnout rate among registered voters across the 83 Michigan counties.

This model demands the most from the data, since it requires sufficient separation between the observations of $y_{ijt-1}$ and $y_{ijt}$ to identify each of their parameters, which is likely difficult within the imputation process since many of the estimates of $y_{ij1}$ will be based on $y_{ijt-1}$.

Table 3 compares the three model specifications in terms of their ability to classify voters and counties by turnout rates. Like the other three models, August turnout predictions perform slightly better than the November turnout estimates. In comparison to Table 1, however, we see that all three panel model predictions have a higher aROC value than the top performing vote lag model. Moreover, there is no evidence of an improvement in performance as the model becomes more complex. The aROC is essentially the same across all three specifications and the dynamic panel estimates of county-level turnout rates has the worst correlation with observed rates. In general, beyond the precinct lag's estimate of county-level turnout rates in November correlating near .9 with observed rates, the basic and precinct lag models perform similarly in their aROC and August county level estimate.

The empirical Bayes predictions show only a slightly better ability to classify high and low turnout voters and counties than the previous vote lag model. But their superior performance is much more evident when we examine absolute rates...
Figure 3. County-level Errors in Turnout Rate (top) and Turnout Count (bottom) Panel Predictions
Table 4. Comparing County-level Accuracy of Panel Model Predictions

<table>
<thead>
<tr>
<th>Model</th>
<th>Election</th>
<th>August</th>
<th>November</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Bias</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td>1.47</td>
<td>0.40</td>
</tr>
<tr>
<td>Precinct Lag</td>
<td></td>
<td>1.82</td>
<td>1.68</td>
</tr>
<tr>
<td>Vote Lag</td>
<td></td>
<td>2.32</td>
<td>-1.30</td>
</tr>
<tr>
<td><em>County-level RMSE</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td>3.69</td>
<td>2.69</td>
</tr>
<tr>
<td>Precinct Lag</td>
<td></td>
<td>3.95</td>
<td>2.65</td>
</tr>
<tr>
<td>Vote Lag</td>
<td></td>
<td>4.54</td>
<td>3.66</td>
</tr>
</tbody>
</table>

Bias and RMSE are expressed in terms percentage point difference in turnout rate among registered voters across the 83 Michigan counties.

of error. Figure 3 displays the county-level rate and count errors using the same scale that was used in Figure 2. Instead of seeing a map bathed in one color, we see the empirical Bayes estimates are much more efficient and unbiased in their 2014 predictions. The top panel shows much more a mix of blue, red, and white, illustrating a symmetric and zero-centered set of errors. The bottom panel uses the same scale as in Figure 2, and it appears almost relatively white, as the errors in predicting the total number of votes within each county is much smaller across the three panel model estimates.

Table 4 measures the bias and efficiency of each county rate prediction, which range from a half to a third of the size of those estimates in Table 2. Moreover, the simplest basic panel model consistently comes out on top. It has the smallest bias in August and November, and a very similar RMSE as the precinct lag model. What Table 4 misses, however, is that the biggest errors in the precinct lag's predictions occur in the more populous Wayne and Oakland county. As seen in the bottom panel of Figure 3, the dynamic and basic panel estimates have much smaller count errors in November, with the basic panel model having the smallest count errors in August.

To further illustrate why these empirical Bayes estimates do so well, Figure 4 shows the county-level association between the average value of \( \hat{u}_i \) from the basic
Figure 4. County-level Averages of $\hat{u}_i$ and Census Characteristics
Figure 5. Non-normality in Observed Distribution of $\hat{u}_i$ (Basic Panel)

panel model and either its logged median household income (top panel) or percent of adults with a bachelors degree (bottom panel). The county average of unobserved tendencies to turn out to vote shows a fairly strong association with both a county’s typical level of income and education.

In summary, voter turnout predictions show far superior performance when accounting for a voter’s prior history of participating. But it seems that the best way to use prior history is as an indicator of the type of voter each person is rather than simply use lagged turnout measures as a predictor of future turnout. Moreover, another apparent advantage of an empirical Bayes approach is that it seems to be less sensitive to specification changes. Estimates and error rates were very similar across specifications and were much less smaller in magnitude.

Does Classifying More than Voters Improve Performance?

Although impressive, a key limitation to classifying voters simply based on their past performance is that we cannot classify voters who lack a recorded history. Predictions for these 13% of registered voters without a history will only be based on what is the average rate of turning out to vote in the forthcoming election cycle for
voters of that age and gender. As shown in Figure 5, many voters receive a zero estimate of $u_i$ since they have no history. We also see that the empirical Bayes estimates of $u_i$ for the other voters fails to follow the assumed normal distribution, as there is bimodal clustering among the ranges -1 to -.3 and .2 to 1.

Perhaps the inclusion of census measures of rates of income and education in each census block would make the unobserved component more normal in shape. But is unlikely that it can make it approximate a unimodal distribution. Another possibility is that this clustering is a function of something broader than a neighborhood's demographic characteristics.

Consider Figure 6, which maps out the sum of each $\hat{u}_i$ estimate by subsections of the Detroit metro area. Although $\hat{u}_i$ is not substantively meaningful on its own, by taking the sum of it within geographic areas we can pinpoint areas where there is a thick concentration of one type of behavior, especially since residential urban density is consistent across this region. Here the red shading indicates sums of $\hat{u}_i$ that add up to be much higher than zero, thereby indicating areas where unobserved forces make turnout typically higher than what is expected based on its age profile alone. Likewise, blue values indicate areas were turnout is much lower than what would be expected simply based on the age of an area's residents.

Detroit residents have much higher rates of poverty and lower rates of education relative to other Michiganders, and we see the effects of that here. The city of Detroit is uniformly blue, whereas the richer suburbs (Grosse Pointe, Royal Oak, Bloomfield Hills, Novi, …) are equally uniform in their reddish tone. Certainly some of these jurisdictions tend to be richer or poorer, but what is striking is the uniformity of these extreme tendencies of being less or more likely to turn out. In comparison, only Dearborn and Warren stand out with a mix of white and color within their boundaries, thus indicating a mix of voters with differing predispositions to turn out. Since Detroit area residents account for such a large portion of Michigan's population, accounting for these jurisdiction-level tendencies with an additional random error component has the potential to inform predictions for voters without a history and make the individual-level error component more unimodal.

I estimated a multi-level logit model and then generated empirical Bayes estimates of unobserved error components at the individual and jurisdiction level. These estimates were generated using the stratified subsampling estimator, where a random subsample of 20 voters were selected from each jurisdiction. Table 5

---

7. The visualization of means is obscured by industrial or commercial sections that contain only a few voters.
presents the performance of predictions. The stars indicate the where the multi-level model improved over the individual-panel models, as presented in Tables 3 and 4.

Estimating unobserved attributes at both the individual and jurisdiction level appears to add only small amounts of predictive performance, and only for the August primary predictions. The aROC, bias, and RMSE are slightly better in August, but the metrics for the November general election are some of the poorest among the empirical Bayes estimates. It is possible that the stratified subsampling estima-
Table 5. Does Accounting for Registration Jurisdiction Improve Performance?

<table>
<thead>
<tr>
<th>Metric</th>
<th>Election</th>
<th>August</th>
<th>November</th>
</tr>
</thead>
<tbody>
<tr>
<td>aROC</td>
<td></td>
<td>.865*</td>
<td>.799</td>
</tr>
<tr>
<td>County-level Correlation</td>
<td></td>
<td>.675</td>
<td>.810</td>
</tr>
<tr>
<td>County-level Bias</td>
<td></td>
<td>0.62*</td>
<td>1.85</td>
</tr>
<tr>
<td>County-level RMSE</td>
<td></td>
<td>3.54*</td>
<td>3.66</td>
</tr>
</tbody>
</table>

* - indicates an improvement over the panel model performance presented in Tables 3 and 4. County-level estimates compare actual turnout rate among registered voters among the 83 Michigan counties.

tion routine contributed to the model’s poor performance. But this seems unlikely since the August primary results show slight improvements. The more likely explanation is that the simpler individual panel model is more robust to errors in making out-of-sample predictions.

Summary and Future Directions

Despite the size of state voter lists and the limits of what they record, these findings show that there are practical ways to use them to accurately forecast voter turnout at both the individual and aggregate level. Both the simulation and application evidence show that a subsampling estimator can do fairly well in generating turnout predictions, as both the aROC and the county-level error rates of the panel models appear to be within a range that are useful for estimating what voters make up electorates within small local jurisdictions. Scholars with limited computational resources can generate accurate parameter estimates by adjusting the size of the subsamples to fit what memory and cores are available to them.

When choosing between alternative modeling frameworks, the Michigan evidence consistently shows that using a voter’s personal history as an indicator for classification purposes performs better than an observed variables approach that predicted future turnout from past behavior. Acknowledging that this evidence is only from comparing performance within one cycle and within one state, it still represents potentially good news in a couple of ways. First, the bias and efficiency of the panel estimators was much less sensitive to the inclusion of different types
of variables. Whereas the bias of the precinct lag and vote lag models changed by over 10 percentage points in the observed variables approaches, it only changed by a range of 3 points across the empirical Bayes predictions. And a second potential benefit of the classification approach is that imputation of past behavior may not be necessary. While its predictions are no different from a basic logit when there is no observable history, at least it can offer predictions without having to rely on imputation.

These findings, however, only represent a first step (and a first cut). There remain many more potential ways to improve upon these estimates. Although including jurisdiction-level effects predictions did not improve predictions, there might be better ways to classify how turnout behavior clusters at the neighborhood or city level in order to predict turnout among the newly registered. For instance, one possibility is to nest a bootstrap imputation routine, that randomly draws a value of $\hat{\mu}_i$ from neighbors or fellow city residents. Another interesting possibility is to use a latent class or mixture model approach that provides further distinction between voters who only turnout in general elections or in presidential election years as opposed to voters who consistently turnout. We can then use observable and address-specific variables to model the mixture of these different types of voters across different cities. It remains to be seen if these models demand too much from the data for it to be reliably accurate in making out-of-sample predictions.

Regardless, accurate turnout predictions can improve election survey analysis and make it easier and cheaper for local candidates to run. There are some initial hassles in obtaining and managing state voter history, but these costs are relatively minor. Moreover, as demonstrated, it seems the costs to developing and estimating predictive models are minor as well.

References


