Partisan Electoral Outcomes and Validated Voter Turnout: A Surprisingly Balanced Electorate

Yair Ghitza

Columbia University
Department of Political Science, Ph.D. candidate
yg2173@columbia.edu

May 1, 2009

Mini-APSA Conference
Columbia University

Abstract

Do turnout levels affect partisan outcomes in elections? Previous work in the literature has come to mixed conclusions. I focus on a single midterm election in Pennsylvania 2006 to investigate. Using a large-scale poll with validated vote data, I run a set of simulations monitoring electoral outcomes under different turnout scenarios, first altering turnout for the entire electorate, then altering turnout for specific demographic subgroups. I come to four main conclusions. First, an individual's past vote history is the strongest predictor of future voting patterns. Second, altering the overall turnout rate of the electorate has minimal impact on electoral outcomes. Third, the electorate is found to be balanced across partisanship and voting likelihood, more so than previously discussed in the literature. Fourth, in part due to this balance, altering the turnout rate of specific subgroups has less of an impact on electoral outcomes than anticipated.

1. Introduction

Do turnout levels influence partisan outcomes in elections? The conventional wisdom among political practitioners and the general public is that higher levels of voter turnout should push electoral outcomes in favor of the Democrats. Because nonvoters are generally associated with low socioeconomic status such as low education and income (Wolfinger & Rosenstone 1978; Leighley & Nagler 2007), their interests should align with the liberal policies of the party to the left.

While this concept seems straightforward, the scholarly literature on the topic comes to mixed conclusions. DeNardo (1980) argues there should be little partisan effect, using a formal model to show that "peripheral" voters are less likely to hold strong partisan leanings. Radcliff (1994) shows a substantial pro-Democrat effect associated with higher turnout, though Erikson (1995) demonstrates Radcliff's findings are due to methodological flaws. Leighley & Nagler (2007) reiterate the demographic similarities between Democrats and nonvoters, but Nagel & McNulty (1996) argue the partisan effect has been insignificant in senatorial and gubernatorial races since 1965. Advanced statistical methods such as simulation and propensity score weighting have been used to address the topic, as well. Citrin, Schickler, & Sides (2003) use simulation methods on Senate races to show that while nonvoters are more Democratic than voters, the lack of close races means that very few election outcomes would have changed had everyone voted. Martinez & Gill (2005) simulate turnout for selected presidential elections, showing that partisan effects have decreased over time. Brunell & DiNardo (2004) use propensity score weighting to show that while nonvoters are slightly more Democratic, no presidential election outcomes since 1952 would have been changed with higher turnout, with the possible exceptions of 1980 and 2000.

Instead of investigating turnout effects across states or across elections, I add to the theoretical discussion by examining turnout effects for one specific election—the Pennsylvania 2006 midterm senatorial election—in great detail. Here I briefly describe the data used for the analysis, research design, and conclusions, after which I will describe each in more detail.

I use two unique datasets to conduct the analysis. The first is a large-scale survey conducted in summer 2006 in Pennsylvania. The size of the survey (N=8,000) and the extent of detailed demographic and partisanship questions are features rarely found in a survey for a single state. The second is the Pennsylvania voter registration file, which I use to append validated vote and registration data to the survey respondents. The voter registration file is a wonderful resource, because it provides individual-level validated vote data for the 2006 election and for numerous past elections. In combination, these datasets provide detailed demographic, registration, and validated turnout information for a large-scale random sample.

I conduct the analysis in four stages. First, I use multivariate logistic regression to assign individual-level probabilities of turning out and supporting each candidate to each survey respondent. With these probabilities, I follow Martinez & Gill (2005) by simulating partisan outcomes on different levels of turnout. I continue with an analysis of the distribution of the electorate in terms of partisanship, voting likelihood, and

demographics. Lastly, I perform additional simulations on particular subgroups in the electorate.

Corresponding to the four stages of the analysis, I come to four main conclusions. First, despite the inclusion of numerous demographic control variables, an individual's past vote history is by far the strongest predictor of future voting patterns. Second, altering the overall turnout rate of the electorate has minimal impact on electoral outcomes. Third, the electorate is found to be balanced across partisanship and voting likelihood, more so than previously discussed in the literature. Regression analysis reveals the demographic makeup of Democratic and Republican voting (and nonvoting) blocs. Fourth, in part due to this balance, altering the turnout rate of specific subgroups has less of an impact on electoral outcomes than anticipated.

2. Data

As the base of the analysis, I use a large-scale survey (N=8,000) of registered voters in Pennsylvania, conducted from July-August 2006 by Lake Research Partners. Given that I intentionally focus the analysis on a single state, Pennsylvania has a number of advantages. Presidential election results have been close in recent years, with Al Gore defeating George W. Bush (R) 51-46% in 2000, John Kerry defeating George W. Bush in 2004 by an even slighter margin of 51-48%, and Barack Obama eventually expanding the Democratic margin to 55-44% in 2008. It is widely considered a "swing" state which consistently garners a substantial amount of attention from both the media and political campaigns themselves, even more so in 2004 and 2006 than in 2008. In addition, the state is relatively diverse, including two urban centers – Philadelphia and Pittsburgh – as well as substantial suburban and rural areas. Indeed, political consultant James Carville once famously described the state as Paoli (a suburb of Philadelphia) and Penn Hills (a suburb of Pittsburgh) with Alabama in between. With the notable exception of a smaller Hispanic population, most important demographic categories are close to national averages, according to the Census.

Regarding the dataset itself, the survey has two distinct advantages over most public opinion surveys for this type of research. The first obvious advantage is the large sample size which is uncommon for a single poll in a single state over a short period of time. The second important advantage is the sampling methodology. Instead of using a random-digit-dial methodology (RDD), the calling sample was pulled directly from the Pennsylvania state voter registration file, i.e. registration-based sampling (RBS). Green and Gerber (2006) show that in gubernatorial and congressional elections in 2002, RBS performed as well or better than RDD in terms of forecasting accuracy. McDonald (2007) assesses the reliability of voter registration files in ten states plus the District of Columbia and finds that the distribution of the electorate on these voter registration files matches expectations from election administration statistics and the Census's Current Population Survey.

I use a voter registration file provided by Catalist, a Washington D.C. based voter registration file vendor who maintains a comprehensive national database of all voting-age individuals in the United States. Green and Gerber (2006) describe some of the advantages of using voter registration files from vendors instead of directly from the

Secretary of State, noting, for example, "the advantage of purchasing registration lists from vendors is that they have often taken the trouble to append to the data set information about the previous elections in which a person has voted" (Green & Gerber 2006, p. 202). Because the survey was conducted using RBS from a high quality voter registration file, I have the flexibility to append a substantial amount of additional information to the survey dataset.

For the purposes of this analysis, the most important piece of information from the voter registration file is each individual's personal vote history, recorded in detail on the file. Vote *choice* is not recorded on the voter registration file due to anonymity of voting in the United States, but whether a person actually turned out is recorded for all general, primary, and special elections. Though this data has been rarely used in political science, it has enormous potential for the study of electoral participation.

Before moving on to describe dependent and input variables more fully, it is important to mention one drawback from using this particular sample. The survey was conducted within a campaign context, and so the intended group of respondents was all registered voters, *except* voters who were a priori assumed to be extremely unlikely to vote. Specifically, the sample universe was drawn from the group of voters who either voted in the 2004 general election, or had registered in 2005 or 2006. This can be viewed as a stricter application of the idea of "deadwood," described by McDonald (2007) as the attempt to keep registration rolls accurate by purging them of people who no longer live (and vote) at their registration address.

While this sampling strategy is useful from a campaign perspective, it is slightly problematic from a scholarly perspective because it induces a higher turnout rate among the survey sample than in the general electorate. Indeed, survey respondents turned out at 75%, compared to 50% as reported on the Pennsylvania Secretary of State's website. As it turns out, this sampling strategy is not as problematic as it first appears. As I describe more fully in Appendix I, this is primarily for two reasons. First, the choice of sampling universe should only increase expected turnout by roughly 5 percentage points, with the remainder of the difference due to survey response bias. In other words, the primary cause of the increased turnout rate is the fact that survey respondents are generally more likely to be voters. While this is still somewhat problematic, this issue is likely to be present in *most* survey sampling strategies, including RDD and other standard techniques. Second, as I show in Appendix I, the demographic distribution of 2006 Pennsylvania voters, as measured in the survey, closely matches the demographic distribution of 2006 Pennsylvania voters as measured by the 2006 Pennsylvania exit poll. Though this does not prove demographic representativeness, it does imply that the survey respondents come close to matching an expected statewide representative sample.

However, because this analysis directly concerns the impact of increasing turnout among unlikely voters, the results of the analysis should be taken with caution. Whatever results are taken from the analysis by necessity exclude unregistered citizens and those who are *extremely* unlikely to vote.

3. Demographic, and Vote History Variables

The first stage of the analysis is a multivariate logistic regression, where I estimate the probability of voting in 2006 in order to use these probabilities in the simulation later on. I begin by estimating 2006 turnout on demographic variables alone, then add vote history variables to obtain a more reliable estimate. Here I describe the dependent variable, demographic variables, and other vote history variables. Unless otherwise noted, demographic variables are taken from the survey.

Dependent variable

- *g2006.voted* is coded as a 1 if the respondent voted in the 2006 general election, and 0 if the respondent did not vote in the election. This is taken from the voter registration file.

Demographic input variables

- *female* is coded 1 for females, 0 for males.
- *age* is a continuous variable.
- *minority* is an indicator variable indicating any race other than white (including "other").
- *single* and *marr.other* (widowed, divorced, or unmarried with partner) are indicator variables, with married as the (excluded) base category.
- *children* is an indicator variable indicating presence of children under the age of 18 at home.
- *educ* is an ordinal variable for education level, ranging from 1 (non-HS graduate) to 5 (post-graduate school).
- *income* is an ordinal variable for household income, ranging from 1 (below $20k) to 7 (over $100k).
- *catholic* and *relig.other* (Jewish, Muslim, Other) are indicator variables for religion, with Protestants as the (excluded) base category.
- *employed.retired* and *employed.other* (part-time, unemployed, homemakers, or students) are indicator variables, with employed full-time as the (excluded) base category.
- *union* is an indicator variable for union membership (retired or active), and *union.hh* is an indicator for having a family member in a union (retired or active). Non-union members are the (excluded) base category.
- *evang* is an indicator variable indicating the respondent is either a fundamentalist or an evangelical Christian.
- *relig.attend* is an ordinal variable for frequency of religious attendance, ranging from 0 (never) to 5 (more than once a week)
- *gunowner* is an indicator variable indicating firearms in the household.
- *army* is an indicator variable for armed services membership, with *army.fam* indicating a family member is in the armed services, and non-membership as the (excluded) base category.

- *pop_ 1mile* is a continuous variable indicating population density. This is computed from the voter registration file by counting the number of people that live within 1 mile of the survey respondent, based on Euclidean distance of latitude/longitude household coordinates.
- *lor* is a continuous variable indicating length of residence in years (from the voter registration file).
- *regyear* is a continuous variable indicating year of registration (from the voter registration file).

I impute missing values for all demographic variables, using linear regression for continuous and ordinal variables and logistic regression for indicator variables. I use only the other demographic variables for each of these imputation regressions (i.e. I do not use vote history or vote preference). Imputation is used in order to utilize the full sample size in the survey and is generally expected to bias all results downwards towards null findings. Indeed, unreported analyses were performed using non-imputed demographic data, and all results were substantively similar, with effect sizes slightly higher.

Vote history variables

- *g2000.voted* is an indicator variable indicating that the respondent voted in the 2000 general election. *g2000.nonreg* is an indicator variable indicating that the respondent was not registered in time to vote in the 2000 general election. The (excluded) base category is therefore people who were registered in 2000 but did not vote in that election.
- Similar variables are used for general election vote history from 2001-2005.

Noticeably absent from the analysis is vote history from primary and special elections. Though the voter registration file includes vote history from these elections, I choose to exclude them from the analysis for two reasons. First, participation in these elections can be viewed as partisan indicators, especially for primary elections. Given my focus on turnout effects on partisan outcomes, it is important to use a turnout measure that is not directly influenced by partisan leanings. Second, the scarcity of primary and special election voters leads to instability in point estimates, sometimes making the coefficients for these variables *negative* and statistically significant. Because there is little theoretical justification for these results, it is likely these are artifacts of small sample size.

4. 2006 Turnout Model

I begin by estimating 2006 general election turnout on the set of demographic variables. I then add previous vote history to the equation. Given the expected similarity between past voting history and 2006 voter turnout, I expect the vote history variables to be the most powerful predictors of 2006 turnout:

Hypothesis 1: past vote history is the most powerful predictor of future electoral participation.

Hypothesis 2: similarly, most relationships between demographics and turnout will be accounted for by past vote history.

These expectations are in line with much of the practical targeting work being done by modern campaigns (see Malchow 2008). Formally, the models are estimated as:

$$P(Y_V=1) = \text{logit}^{-1}(\boldsymbol{\beta}_{D1}\boldsymbol{X}_D), \quad (1)$$

$$P(Y_V=1) = \text{logit}^{-1}(\boldsymbol{\beta}_{D2}\boldsymbol{X}_D + \boldsymbol{\beta}_V\boldsymbol{X}_V), \quad (2)$$

where Y_V is the dependent variable (indicating whether the respondent voted), $\boldsymbol{X}_D$ is a matrix of demographic variables (including a column of 1s), $\boldsymbol{\beta}_{D1}$ and $\boldsymbol{\beta}_{D2}$ are vectors of parameter estimates (with the first term the constant term), $\boldsymbol{X}_V$ is a matrix of vote history variables, and $\boldsymbol{\beta}_V$ is another vector of parameter estimates.

Figure 1 displays the model estimates. The left-hand plot shows variable coefficients from model (1), with variables significant at the 95% level shown in red. In all regression analyses, continuous and ordinal variables are standardized by subtracting the mean and dividing by two standard deviations, as recommended by Gelman & Hill (2007), allowing for easy comparison of effect size across variables. There are no big surprises in the demographic coefficients. *Age* and *education* are positively correlated with voting, following expectations from the literature. Similarly, *regyear* is negative, indicating respondents who have been registered for longer tend to vote more. Also positive is *union* membership, which is unsurprising considering the amount of resources devoted to mobilization from unions. *relig.attend* is positive, indicating stability in the community, perhaps along with mobilization efforts from Republican organizations. *Length of residence* is positive, also indicating stability in the community. *Minorities* and *non-married* people are less likely to vote, which is also unsurprising. Slightly surprising are the null findings for *children* and *income*, though *income* is less surprising considering the relatively high correlation with *education* (0.49).

The middle plot adds vote history input variables, as indicated in model (2). Strong evidence supporting Hypothesis 1 is immediately apparent. First, a likelihood ratio test (LR-statistic=1009; d.f.=13) indicates the model is a much better fit than model (1). More importantly, the vote history variables are clearly the most powerful predictors of 2006 turnout. *g2005.voted* is the most powerful predictor, which is unsurprising considering it is the most recent off-year election. Also powerful are *g2003.voted*, *g2002.voted*, and, to a lesser extent, *g2001.voted*.

The null finding and large standard errors for *g2004.voted* are statistical artifacts from the sampling procedure. Recall the sample was intended to be people who voted in 2004 or registered since then. If this were the case, either *g2004.voted* or *g2004.nonreg* would have to be removed from the analysis (both variables included would imply no base term). As it turns out, 82 respondents are coded as having been registered in 2004 but not voted. The exact cause of this small discrepancy is unknown,

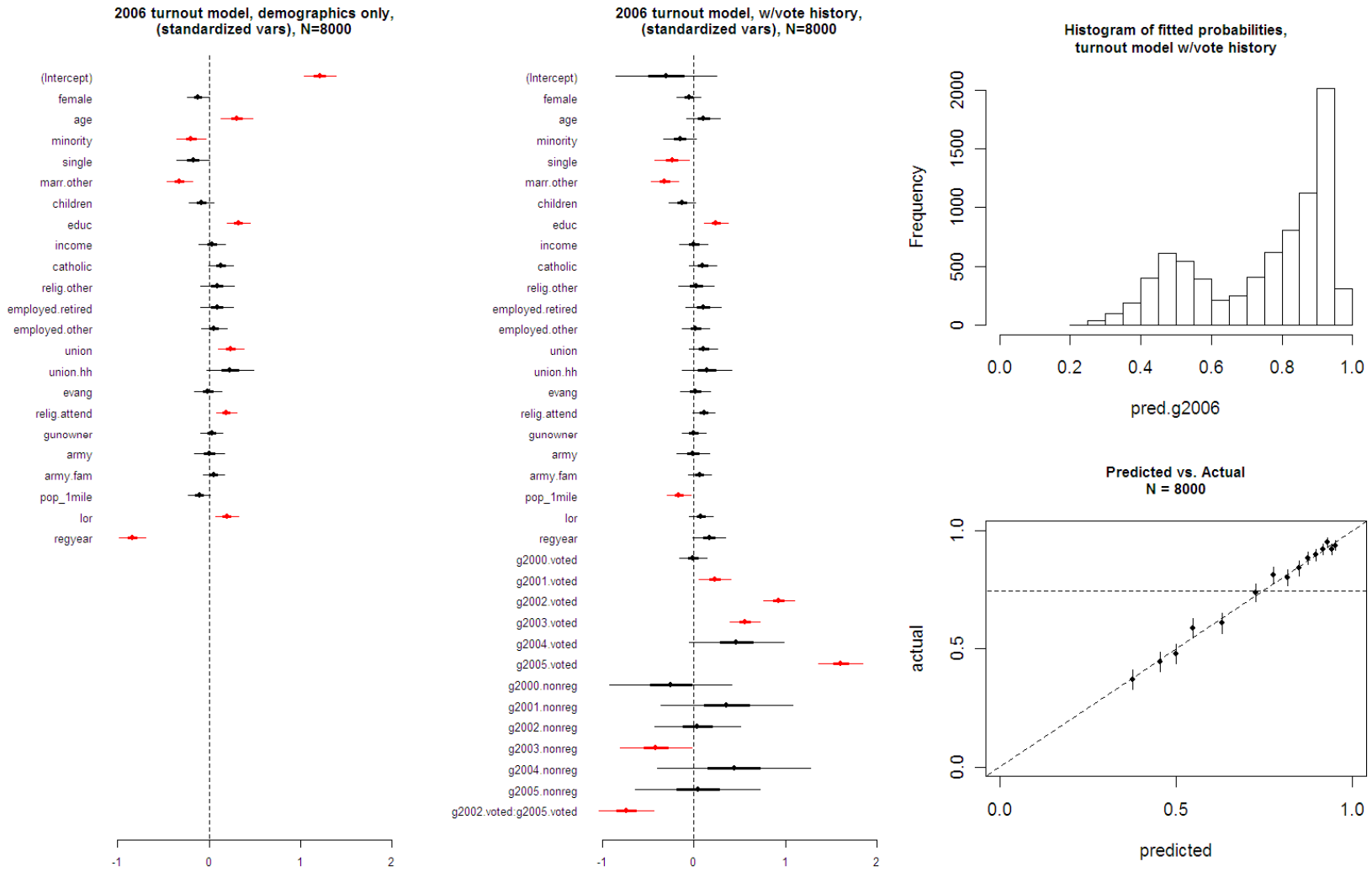

Figure 1 *Turnout model, with demographics and vote history. The left-hand plot displays variable coefficients for model (1) with demographics only, along with 50% and 95% error bars. Red coefficients indicate statistical significance at the 95% level. Most demographic coefficients are in the expected direction. The middle plot adds vote history input variables, as indicated in model (2). The strongest coefficients are those for the vote history variables, with most demographic coefficients shrinking toward zero. The top-right plot shows the distribution of fitted values from model (2), while the bottom-right plot shows the predicted probability against the actual probability (calculated in 15 bins) with data level error. The close fit around the diagonal reference line indicates a strong model fit.*

but it is likely due to the fact that the original sample was pulled from an earlier version of the voter registration file, and so vote history or registration dates were likely updated in the more recent version.

Indicator variables for non-registration during a particular past election are generally non-significant, all with large standard errors. The final term is an interaction between *g2002.voted* and *g2005.voted*. Given that these two variablesare the most highly significant, the interaction term helps normalize estimates to avoid overestimating the likelihood of past voters to vote again, and vice versa.

Moving to Hypothesis 2, notice that, with the exception of *education* and *non-married* people, all other variables are reduced to be indistinguishable from zero. *pop_ 1mile* is now significant, though it is little changed in magnitude from model (1). Though all other demographic variables are non-significant, it would be incorrect to conclude that these other variables are not correlated with 2006 turnout. Rather, as shown in model (1), the standard demographic relationships still hold up with this data. Further, past vote history can be estimated using demographic variables alone. However, these results do provide evidence for Hypothesis 2: most of the variation in turnout described by demographics can be accounted for using past vote history alone.

The top-right plot in Figure 1 displays the distribution of fitted values from model (2), with the distribution skewing upward due to the high turnout level of the survey respondents. The bottom-right plot shows the predicted probability against the actual probability (calculated in 15 bins) with data level error. The data level error is due to uncertainty on the parameter estimates, which causes uncertainty in the predicted probability. Notice the points in this plot vary closely around the diagonal reference line, indicating a strong model fit. This is a good sign, because I will be using these predicted probabilities to simulate different levels of turnout later in the analysis.

5. Senate Vote Choice Model

I now estimate vote choice using demographic variables and additional partisanship measures from the survey. Given the high correlation between vote choice and partisanship, these input variables help obtain more reliable estimates for probabilities of vote choice. Here I describe the dependent and other input variables to the model. Unless otherwise noted, all of the following variables come from the survey:

Dependent variable

- *dep.sen.orig* is coded as a 1 if the respondent indicated he/she would support the Democratic candidate Bob Casey Jr. and 0 for supporters of the Republican candidate Rick Santorum. "Leaners" were included in this analysis, but respondents who answered they were undecided were removed. The full range of responses was: strong Santorum, not strong Santorum, lean Santorum, undecided, lean Casey, not strong Casey, and strong Casey.

Partisan input variables

- *ID.dem* is an ordinal variable indicating level of Democratic party identification. This is coded from a standard 7-pt party ID question—Strong D, Not strong D,

Lean D, Independent, Lean R, Not strong R, and Strong R—and these responses are coded from 3 to -3, respectively.

- *vote.dem.gov* and *vote.dem.house* are ordinal variables indicating support for the gubernatorial and U.S. House candidates, respectively. These are coded from 3 (indicating highest Democrat support) to -3 (indicating highest Republican support), where possible responses are analogous to the Senate vote question.
- *reg.dem* is an ordinal variable indicating party registration status, taken from the voter registration file. Registered Democrats are coded as a 1, registered Republicans are coded as a 0, and people registered as Independents or some other party are coded as 0.5.
- *diff_ 1mile.p* is a continuous variable indicating localized partisanship, computed from the voter registration file. This is computed by counting the number of registered Democrats and Republicans that live within 1 mile of the survey respondent, based on Euclidean distance of latitude/longitude household coordinates, and computing the percent Democrat advantage for each survey respondent. The measure therefore spans from -1 (indicating the area within 1 mile of the respondent is 100% Republican) to 0 (indicating a 50/50 split), to 1 (indicating 100% Democratic).

I estimate the following model:

$$P(Y_C=1) = \text{logit}^{-1}(\boldsymbol{\beta}_{D3}\boldsymbol{X}_D + \boldsymbol{\beta}_P\boldsymbol{X}_P), \qquad (3)$$

where Y_C is the dependent variable (indicating vote *choice*), $\boldsymbol{X}_D$ is a matrix of demographic variables (including a column of 1s), $\boldsymbol{\beta}_{D3}$ is a vector of parameter estimates (with the first term the constant term), $\boldsymbol{X}_P$ is a matrix of partisanship variables, and $\boldsymbol{\beta}_P$ is another vector of parameter estimates.

Figure 2 displays the model estimates. Again, the left-hand plot shows variable coefficients from the model, with variables significant at the 95% level shown in red. As expected, the partisanship measures are by far the strongest predictors of senate vote choice. All partisan measures (with the exception of *diff_ 1mile.p*) are significant; I also include interaction terms between the most highly significant variables to dampen the extreme likelihoods of strong partisans. Besides the partisan variables, *education*, *income*, and *union* membership are all slight indicators of Democratic vote choice, while being a *Catholic*, *evangelical*, or high *religious attender* are all slight indicators of Republican vote choice.

Again, the top-right plot in Figure 2 displays the distribution of fitted values from the model. The bipolar distribution indicates that most of the respondents are identified as likely Casey supporters or likely Santorum supporters. The bottom-right plot again shows the predicted probability against actual probability (calculated in 15 bins) with data level error. Again, the close fit to the diagonal reference line indicates a strong model fit.

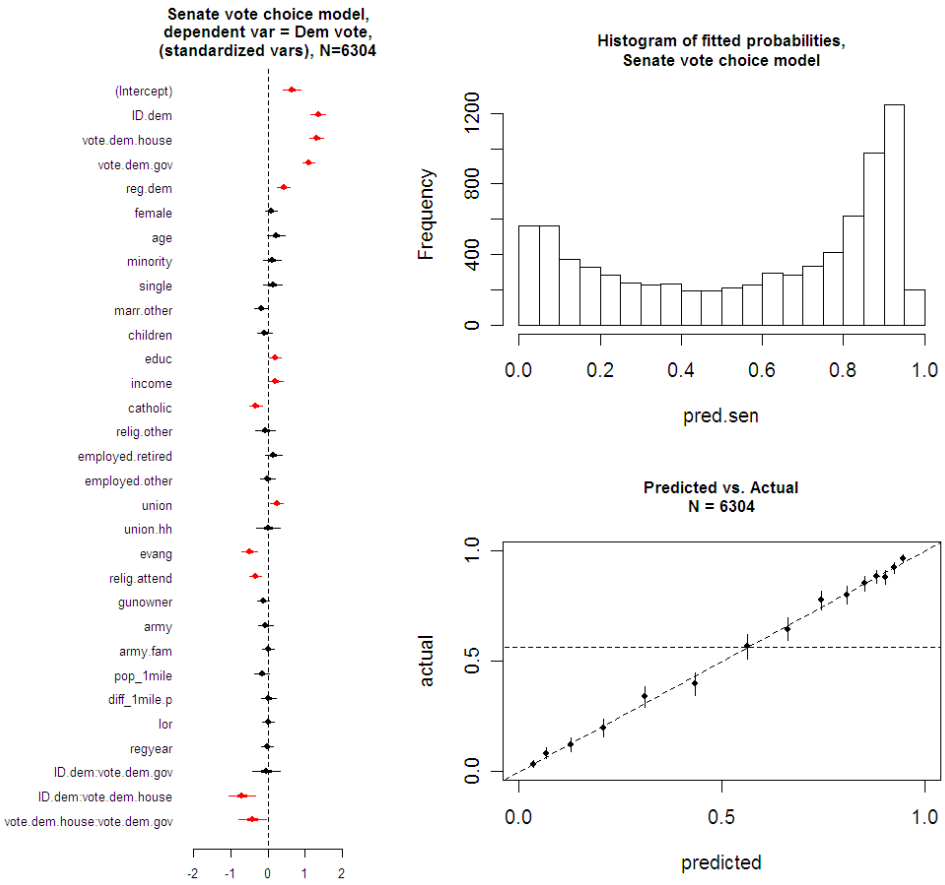

Figure 2 *Senate vote choice model. The left-hand plot displays variable coefficients for model (3), along with 50% and 95% error bars. Red coefficients indicate statistical significance at the 95% level. As expected, partisan indicators are the most significant variables. The top-right plot shows the distribution of fitted values, and the bottom-right plot shows the predicted probability against the actual probability (calculated in 15 bins) with data level error. The bipolar distribution of fitted values and close fit around the diagonal reference line indicate a strong model fit.*

---

6. Simulating Partisan Outcomes Under Various Turnout Levels

Now that the turnout and vote choice models are completed, I can apply the models to create predicted probabilities of turning out and supporting the Democratic candidate for each respondent in the survey. In this section, I use these probabilities to simulate different levels of turnout and observe how these changes affect partisan outcomes in the election. This is similar to the simulation method conducted by Martinez & Gill (2005).

My simulation method works as follows. In order to achieve a certain level of turnout, I change the intercept of the turnout model by adding some constant δ, which

is positive in the case of simulating higher turnout and negative in the case of simulating lower turnout. This creates new predicted probabilities of voting for each survey respondent, denoted V. Remember each respondent also has a predicted probability of supporting the Democratic candidate, denoted C for vote choice, which does not change during the simulation. For each turnout level, I simulate a fake election by drawing from the binomial twice—first using V to indicate whether the respondent "votes" in the election, second using C to indicate whether the respondent will support the Democratic or Republican candidate, conditional on voting. I perform this simulation a certain number of times—referred to as *iterations*—for each turnout level, and I report the mean election outcome from all iterations for each turnout level. By simulating turnout levels going from 0% turnout to 100% turnout, I examine the partisan effects of changing voter turnout levels to all possible levels of turnout. Formally, the process can be described as follows:

$$V = \text{logit}^{-1}(\boldsymbol{\beta_{D2}X_D} + \boldsymbol{\beta_V X_V} + \delta)$$
$$C = \text{logit}^{-1}(\boldsymbol{\beta_{D3}X_D} + \boldsymbol{\beta_P X_P}),$$

where V become the fitted values using model (2) along with the δ term, and C become the fitted values from model (3). Then, for each iteration at this turnout level:

$$\tilde{V}_i \sim \text{binom}(1, V_i)$$
$$\tilde{C}_i \sim \text{binom}(1, C_i),$$

where $\tilde{V}_i = 1$ if the respondent votes in the simulated election, 0 otherwise, and $\tilde{C}_i = 1$ if the respondent supports the Democratic candidate in the simulated election. These are calculated for all N respondents in the survey. Then:

$$\widetilde{D}_{total} = \sum_{i=1}^{N} \tilde{V}_i \,, \forall\, \tilde{C}_i = 1$$
$$\tilde{R}_{total} = \sum_{i=1}^{N} \tilde{V}_i \,, \forall\, \tilde{C}_i = 0,$$

where $\widetilde{D}_{total}$ becomes the total number of Democratic votes, and $\tilde{R}_{total}$ becomes the total number of Republican votes, for the simulated election. With the total votes calculated:

$$\tilde{V}_{total} = \frac{\widetilde{D}_{total} + \tilde{R}_{total}}{N}$$
$$\tilde{O}_{total} = \frac{\widetilde{D}_{total}}{\widetilde{D}_{total} + \tilde{R}_{total}},$$

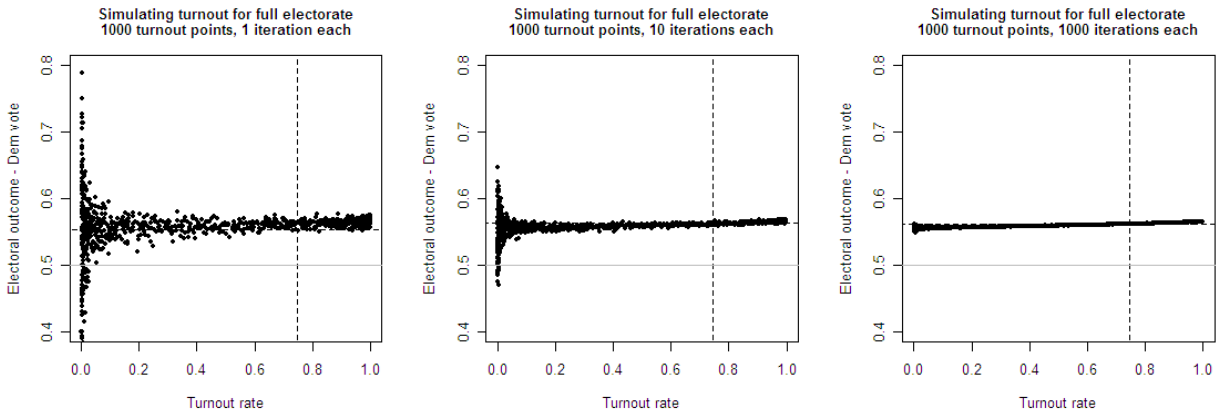

Figure 3 *Simulating different levels of turnout among the entire electorate. The left-hand plot simulates 1000 turnout levels, roughly corresponding to [0.0%, 0.1%, ..., 99.9%, 100.0%], with 1 iteration per turnout level. The x-axis represents the different simulated turnout levels, and the y-axis represents the mean election outcome across all iterations of this turnout level. The dotted lines represent the simulation results at the standard turnout level, i.e. 75%. The middle plot again simulates 1000 turnout levels, this time with 10 iterations per turnout level. The right-most plot moves to 1000 iterations per turnout level.*

---

where $\tilde{V}_{total}$ is the final turnout for this iteration, and $\tilde{O}_{total}$ is the final election outcome for this iteration, i.e. the percentage of the vote that goes to the Democrat. Finally, for each turnout level, I take the mean of all iterations and graph the results.

The simulation results are shown in Figure 3. The left-hand plot simulates 1000 turnout levels, roughly corresponding to [0.0%, 0.1%, ..., 99.9%, 100.0%], with 1 iteration per turnout level. The x-axis represents the different simulated turnout levels, and the y-axis represents the mean election outcome across all iterations of this turnout level. For the left-most plot, because there is only 1 iteration per turnout level, the y-axis simply represents the results from the single iteration. The dotted lines show the simulation results at the standard turnout level, i.e. 75%.

From the left-most plot, a few trends are immediately apparent. First, the trend line is surprisingly flat. Indeed, moving from 0% turnout to 100% turnout yields less than a 1% change in expected partisan outcome! This surprising result suggests that different levels of turnout will have virtually no impact on election outcomes. Second, the variation in election outcome increases dramatically at extremely low turnout levels. This is simply a function of sample size—when less than 1% of the survey votes, a sample size of 80 voters will yield high variation in election outcomes. This artifact of the simulation is resolved in the other two plots in Figure 3. Moving to the middle plot, I again simulate 1000 turnout levels, this time with 10 iterations per turnout level. The trend line stays surprisingly flat, and the variation at lower turnout levels decreases. Finally, moving to the right-most plot, I again simulate 1000 turnout levels, this time

with 1000 iterations per turnout level. The trend line stays flat, and the variation at low turnout levels virtually disappears.

The results of the simulation are striking and surprising. Instead of confirming the conventional wisdom that higher turnout leads to higher vote share for Democrats, these results indicate turnout levels have no effect on partisan outcomes. Why might this be the case? The remainder of this paper attempts to answer this question.

7. Distribution of the Electorate

The results from Figure 3 are surprising at first, but a deeper consideration of exactly how the simulation works reveals clues as to why the trend line is so flat. Recall that in order to simulate different levels of turnout, I add some constant term δ to the intercept for *all* survey respondents. This implies that for each turnout level, though each person's probability of voting V_i changes, the rank-ordering of turnout percentages stays the same. In other words, the survey respondent who is most likely to turn out in the real election will be the most likely to turn out for *all simulated elections*. This is not a problem for the "high turnout" simulations, where all survey respondents vote. However, a "low turnout" simulation is peculiar, because it will in fact most often represent the vote choice distribution among the most highly likely voters.

With this concept in mind, the simulation results are no longer as surprising, and in fact they are quite revealing. The "low turnout" simulation indicates the vote choice preferences of the most highly likely voters—a reasonable simulation given that these people would be most likely to remain in the electorate at such low voting levels—and as the turnout level rises, this simulates incrementally adding people into the electorate based on their relative likelihood of voting. In this sense, the results suggest that the electorate has a partisan balance at all turnout levels. In other words, highly likely voters should be evenly distributed along partisan lines, as are medium likely voters, as are unlikely voters.

> *Hypothesis 3*: the electorate has a partisan balance at all turnout levels. Highly likely voters are evenly distributed along partisan lines, as are medium likely voters, as are unlikely voters.

Note that, for the remainder of this section, I use the phrases "partisan" and "vote choice" interchangeably for ease of exposition. Despite the imprecision of this lexical decision, vote choice and partisanship are closely enough linked in theory as to warrant the gains in expositional clarity.

Due to the detailed nature of this dataset, it is easy to examine the partisan balance of the electorate. Figure 4 displays the distribution of survey respondents by candidate choice and voting likelihood. First notice the 3x3 mosaic plot on the left. The x-axis corresponds to Senate vote choice, as measured directly from the survey, and the y-axis corresponds to the top third, middle third, and bottom third of voting likelihood, as predicted by the turnout model. Each table entry indicates the percentage of survey respondents that fall within that group – for example, 11% of survey respondents are Republican supporters who are highly likely to vote.

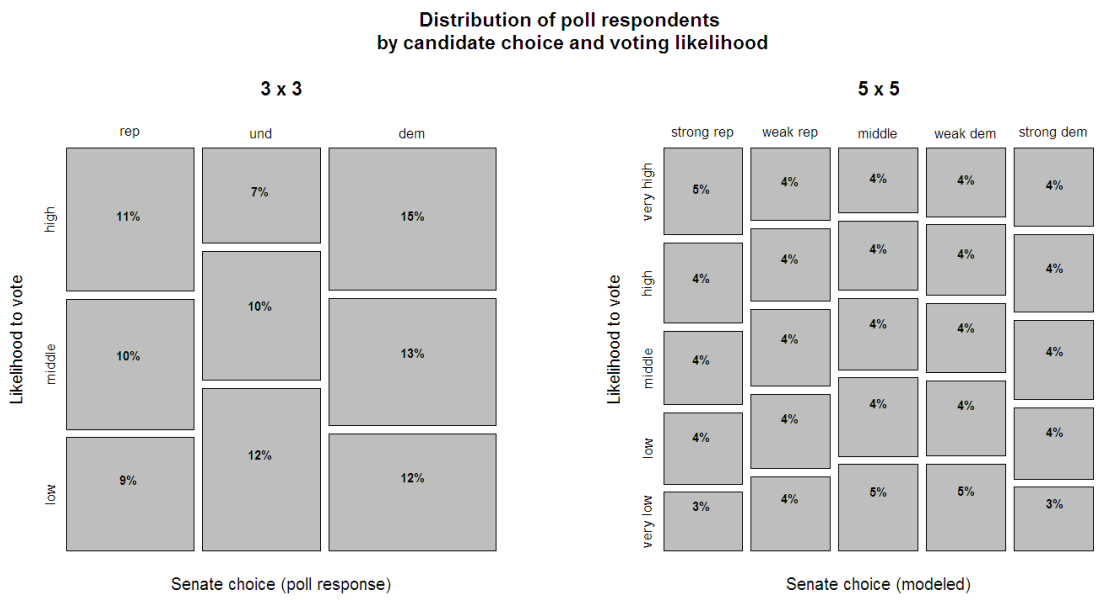

Figure 4 *Distribution of poll respondents by candidate choice and voting likelihood. The left-hand plot distributes survey respondents into a 3x3 grid. The x-axis corresponds to Senate vote choice, as measured directly from the survey, and the y-axis corresponds to the top third, middle third, and bottom third of voting likelihood, as predicted by the turnout model. Each table entry indicates the percentage of survey respondents that fall within that group – for example, 11% of survey respondents are Republican supporters who are highly likely to vote. The right-hand plot splits the electorate into a 5x5 grid, where the x- and y-axes are computed using predicted probabilities from the vote choice and turnout models, respectively.*

The partisan balance across all turnout levels is immediately apparent in Figure 4, as indicated by the similar shape of the Democrat and Republican columns as compared to the middle Undecided column. In numerical terms, notice that among likely voters, the partisan split is 58-42% in favor of the Democrats (58% is calculated as 15 / (11 + 15), where the numerator is the size of the *high-dem* group, and the denominator is *high-dem* + *high-rep*). For the middle and low likelihood to vote groups, the partisan split is 57-43% and 57-43%, respectively. These partisan splits are exactly in line with the election results from Figure 3, which ranged from 56-58% Democratic vote.

This plot also indicates that partisans are more likely to vote than non-partisans. Among partisans, the high-medium-low likelihood to vote split is 37-33-30%, while this split is 24-34-41% for Undecideds. The right-hand plot splits the electorate into a 5x5 grid, where the x- and y-axes are computed using predicted probabilities from the vote choice and turnout models, respectively. The same distributional balance exists while viewing the sample using these finer divisions.

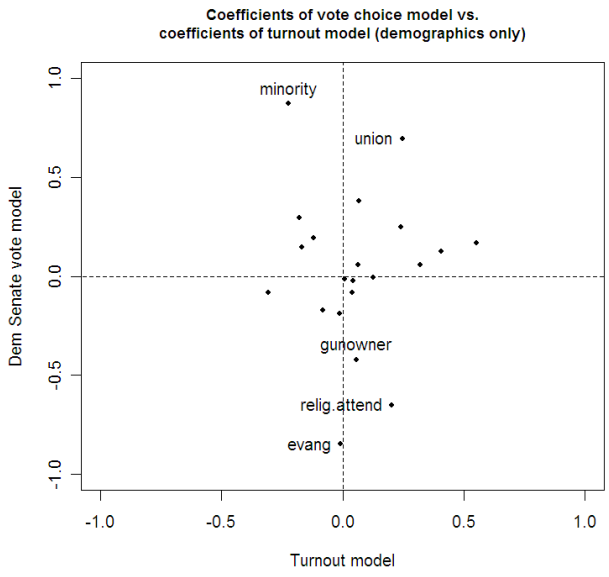

Figure 5 *Demographic distribution of poll respondents by vote choice and turnout likelihood. Each point is a coefficient for a demographic variable. The x-axis is the coefficient for a turnout model using demographic variables only, i.e. model (1) from earlier in the paper. The y-axis is the coefficient for a demographic vote choice model, i.e. a model similar to the earlier vote choice model, but using only demographic variables as input.*

These results lend support to Hypothesis 3, indicating the electorate has a partisan balance at all turnout levels. While it is interesting to note this level of balance, it is unsatisfying to simply view the electorate in terms of partisanship and voting likelihood alone. What *demographic* trends account for these somewhat surprising results?

8. Demographic Analysis

If the electorate is truly balanced as I claim in the previous section, how does the electorate break down demographically? Given the known demographic differences in the makeup of the parties and the different policy platforms offered by the parties, I expect a highly likely voter that is a Democrat to be demographically different from a highly likely voter that is a Republican. Similarly, an unlikely voter who is a Democrat should be different from an unlikely voter that is a Republican.

One way of getting a sense of the *demographic* distribution of the electorate is shown in Figure 5. For this graph, I estimated two logistic regression models—a Senate vote choice model and a turnout model—using only demographic variables as inputs. In other words, the turnout model is the same as model (1) which was graphed in Figure 1, and the Senate vote choice model is similar to the earlier model, except it uses only demographic variables (the full model results are not reported to save space). I remove

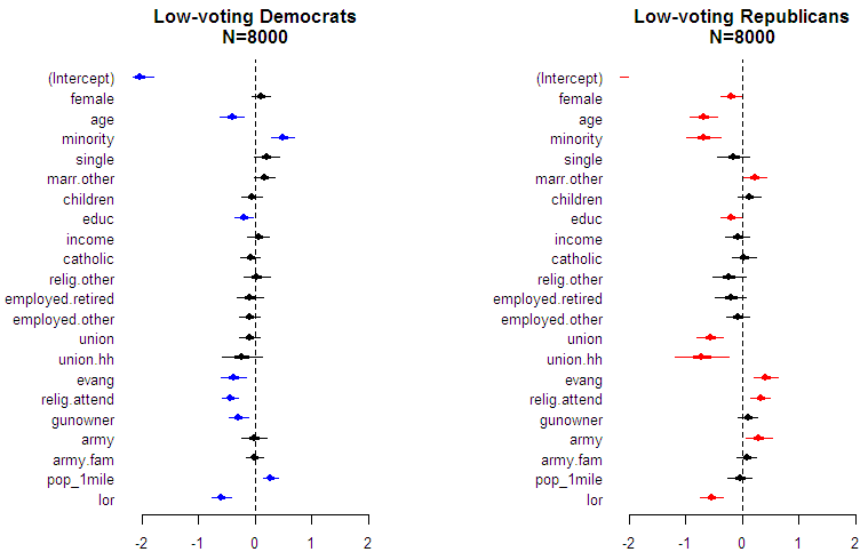

Figure 6 *Demographic characteristics of low-voting Democrats (on the left) and Republicans (on the right). The left-hand plot describes the familiar group of unlikely Democratic voters with low socioeconomic status. The right-hand plot describes a less familiar group: low-voting Republicans who are predominantly white, low educated, and religious.*

---

registration year from the set of input variables because this is more as a political variable than a demographic variable.

In Figure 5, each point is a coefficient for a demographic variable. The location on the x-axis represents the coefficient for the demographic turnout model, and the y-axis represents the coefficient for the demographic vote choice model. For example, the coefficients on *minority* indicate that minorities are less likely to vote and more likely to support the Democratic candidate. Though this single graph does not tell the whole story, it is suggestive of a *demographic* balance in the electorate that may help explain the partisan balance explored in the previous section. Notice the labeled points: *union* members are likely to vote and be Democrats. The mirror image of them in terms of partisanship is *religious attenders*, who are likely to vote and be Republicans. *Minorities* are likely to be Democrats but are unlikely to vote, and, though an exact counterpart on the Republican side is not apparent, it seems *evangelicals* and *gun owners* who are *not* high religious attenders may fit the bill. This final group is not as clear because of the difficulty in interpreting regression coefficients in this way.

I now move to a slightly more formal analysis, where I investigate the demographic characteristics of specific partisan/turnout groups explicitly. Specifically, I estimate a series of four logistic regression models, each model identifying one of the "corner" groups in the left-most plot of Figure 4. The first model, for example, is specified as:

$$P(Y_{DL}=1) = \text{logit}^{-1}(\beta_{DL}X_D), \qquad (4)$$

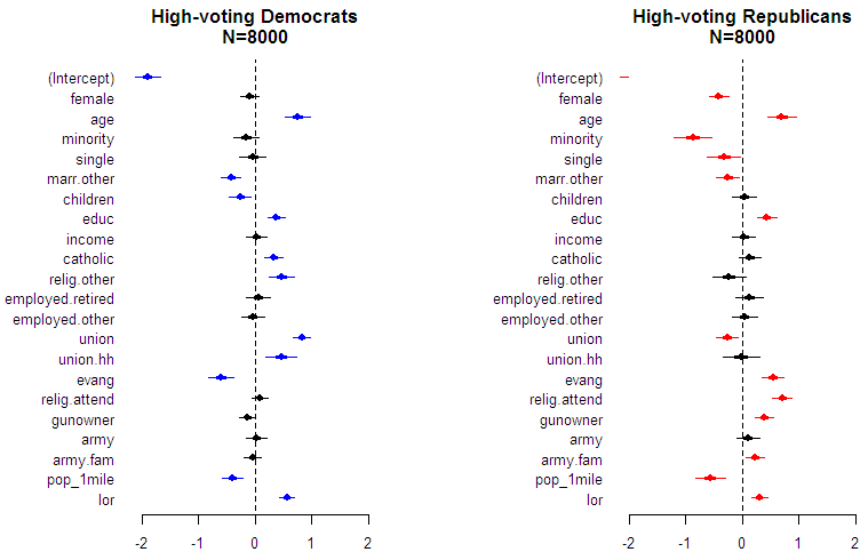

Figure 7 *Demographic characteristics of high-voting Democrats (on the left) and Republicans (on the right).*

---

where Y_{DL} is the dependent variable indicating the respondent is a *D*emocrat with a *L*ow likelihood of voting (coded as 1 for this group and 0 for the remainder of the sample), $\boldsymbol{X_D}$ is a matrix of demographic variables (including a column of 1s), and $\boldsymbol{\beta_{DL}}$ is the vector of parameter estimates for this first model. I also estimate three other models, where the dependent variables are Y_{RL} (for *R*epublicans with a *L*ow likelihood of voting), Y_{DH} (*D*emocrats with a *H*igh likelihood of voting), and Y_{RH} (*R*epublicans with a *H*igh likelihood of voting).

Figure 6 shows the results for the first two models. The left-hand plot shows model coefficients for the model identifying low-voting Democrats. This is a familiar group in the political science literature, comprised of citizens with low socioeconomic status who are unlikely to vote—younger, minorities, low education, urban, not established in the community (as indicated by the negative coefficient for *length of residence*), and non-religious. Under the conventional wisdom, this is the group that will tip elections in favor of the Democrats if there were higher turnout.

The right-hand plot, however, paints a picture of a less familiar group—low-voting Republicans. They are younger, white, low education, religious, also not as established in the community. Members of the armed services are more likely to be in this group, while union members are less likely. Though not as familiar as the previous group, this group of religious, predominantly white, lower socioeconomic citizens who are both nonvoters and likely to be Republicans, seems to be a class of voters who go somewhat unrecognized in the existing literature.

Figure 7 shows the results of the next two models, identifying Democrat and Republican high-likelihood voters, respectively. On the left, high-voting Democrats are found to be older and more established in the community, living in rural areas. They are more educated, non-Protestant, and not religious. Union members are also likely to

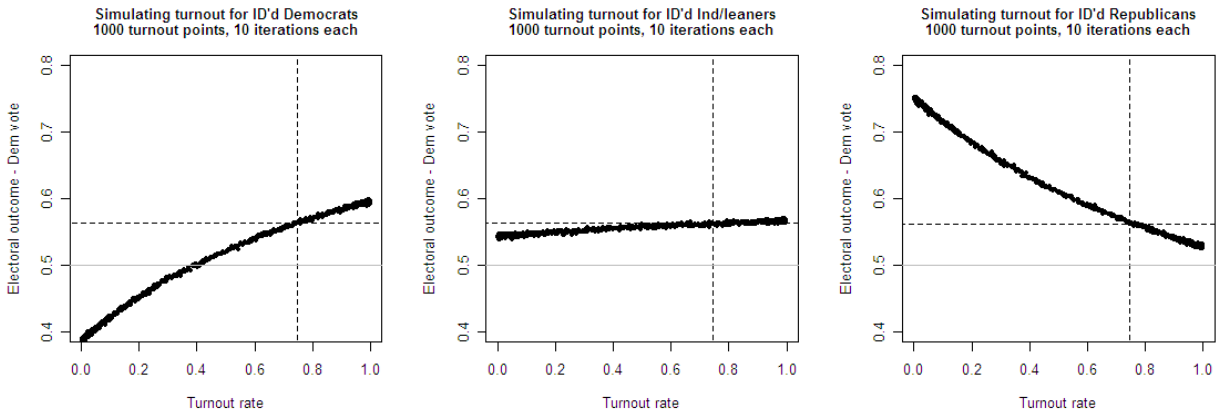

Figure 8 *Simulating different levels of turnout among partisans. Partisanship is defined by the party ID question on the survey, with Democrats on the left, Independents/Leaners in the middle, and Republicans on the right. Removing* all *ID'd Democrats or Republicans from the voting universe changes the electoral outcome by only 20-25 points.*

---

fall into this group, again unsurprising considering the amount of resources put into mobilization by unions. On the right, high-voting Republicans are also found to be older, more established in the community, highly educated, and living in rural areas. Unlike high-voting Democrats, though, they are white, married (the excluded base term), highly religious, and more likely to be gun owners. While these results are not particularly surprising, they describe the electorate in more detail and help confirm the expected demographic characteristics of these groups of voters.

The identification and demographic characterization of these groups is an important contribution to the understanding of partisan outcomes under different levels of turnout. The classical argument that high levels of turnout should yield favorable outcomes for Democrats builds on the known demographic similarities between unlikely voters and Democrats. Through this analysis, I have confirmed the existence and expected demographic characteristics of this low-voting Democratic group, but I have also identified a group of low-voting Republicans who are largely ignored in the literature. It is the existence of this group and the overall balance in the electorate that drives the simulation results presented earlier.

9. Additional Simulations

The simulation framework can be extended to examine turnout scenarios that are more complex than a simple uniform increase or decrease in expected turnout rate across the entire electorate. In this section, I alter the turnout rate for specific demographic subgroups, while holding other subgroups at their actual turnout levels.

In a sense, these simulations should be considered more realistic than the earlier simulations. Most get-out-the-vote (GOTV) or mobilization campaigns are targeted

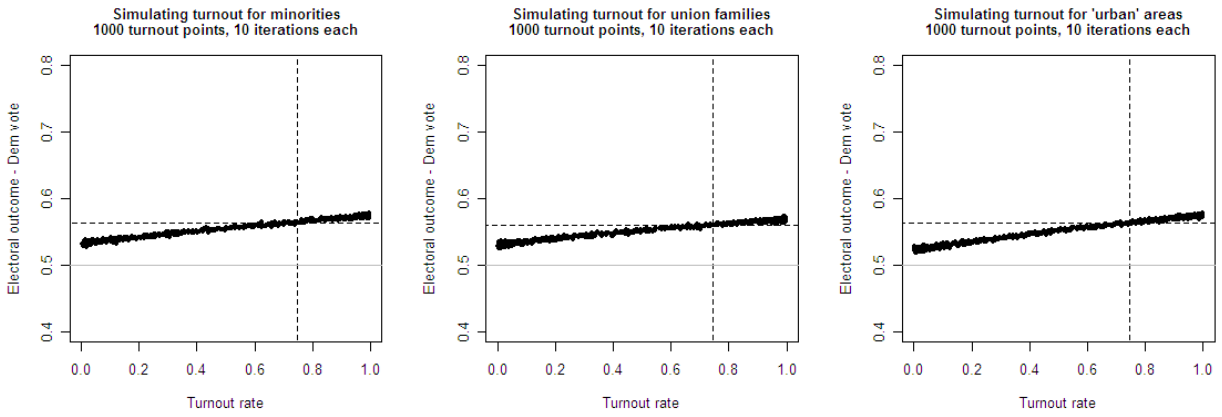

Figure 9 *Simulating different levels of turnout among "Democratic" subgroups: minorities, union families, and residents of "urban" areas. For minorities, the small magnitude of the effect can be explained by the small percentage of minorities in the sample (16%) and the partisan split within this subgroup (76-24% in favor of Casey). Similar numbers can be shown for the other groups.*

---

towards specific demographic or partisan subgroups, rather than towards the entire electorate. In other words, the Democratic Party will try to mobilize likely *Democratic* voters, rather than expend resources on mobilizing Independents or, even less so, Republicans.

In Figure 8, I simulate different levels of turnout for the most obvious group – partisans as identified through the party ID question. The left-most graph simulates the turnout level for ID'd Democrats with 1000 turnout levels and 10 iterations per turnout level. The results of this simulation are again surprising. As expected, there is a clear correlation between simulated Democratic turnout and the partisan outcome, where the partisan outcome changes 20-25 points based on the level of Democratic turnout. However, notice Bob Casey, the Democratic candidate, would still get 40% of the vote even if no Democrats were voting!

Though this result is quite shocking at first, some simple math explains it. ID'd Independents account for 23% of the sample, and they support Casey 65-35% (excluding Undecideds). Similarly, ID'd Republicans are 37% of the sample, and they support Casey 18% of the time. Though we tend to think of partisanship as the ultimate indicator of vote choice, in reality the two are not perfectly correlated. This imperfect correlation and the substantial proportion of Independents in the electorate lead quite naturally to this result.

The middle plot in Figure 8 simulates turnout levels for ID'd Independents/ Leaners, and the right-hand plot simulates turnout levels for ID'd Republicans. The Republican plot follows a similar logic to the Democratic plot, and the middle plot indicates that a higher turnout level for Independents would be slightly advantageous to Casey. However, this middle plot should be taken with caution, as it is clearly driven

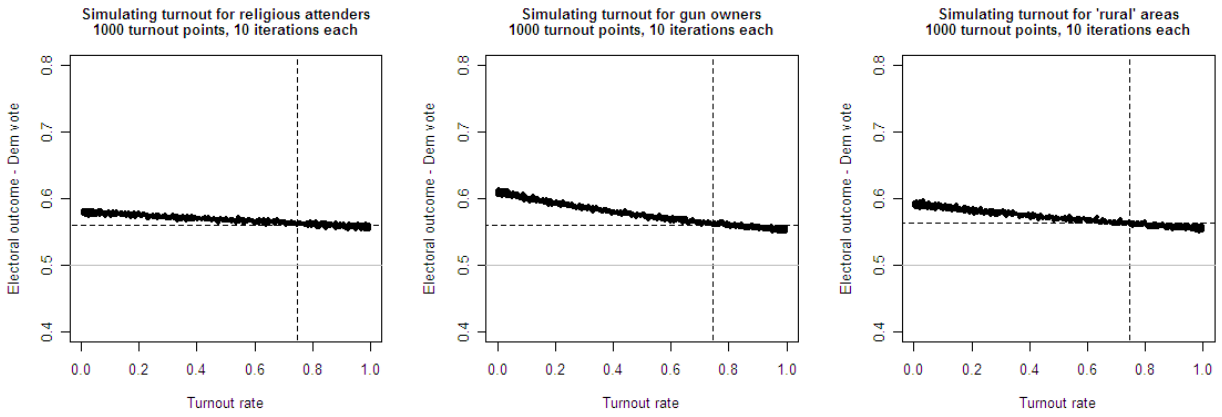

Figure 10 *Simulating different levels of turnout among "Republican" subgroups: religious attenders, gun owners, and residents of "rural" areas. The magnitude of the turnout effect is similar to Figure 9.*

---

by Casey's advantage among Independents, an advantage which may not hold for Democratic candidates in other elections.

Figure 9 simulates turnout levels for three demographic subgroups I consider "Democratic" based on the earlier demographic analyses: minorities, union families (including household members), and residents of "urban" areas, defined as the top 25% of respondents in terms of population density. All of these graphs display a slightly positive trend. Though conventional wisdom would suggest that changing the turnout level among minorities would have a substantial impact on electoral outcomes, the small percentage of minorities in the sample (16%) and the partisan split within this subgroup (76-24% in favor of Casey), lead to the lower than expected slope of the trend line.

Figure 10 shows analogous simulations for "Republican" groups – religious attenders, gun owners, and residents of "rural" areas, defined as the bottom 25% of respondents in terms of population density. The magnitude of the turnout effect for these groups is similarly small.

10. Conclusion

Do turnout levels affect partisan outcomes in elections? While previous work has come to mixed conclusions, I focus on a single election – the Pennsylvania Senate campaign in 2006 – and use a highly detailed data source to investigate. In so doing, I come to a series of conclusions. First, an individual's past vote history is the strongest predictor of future voting patterns, demonstrating the potential research value of voter registration file data in understanding electoral participation in greater detail.

More to the point, I find that altering the overall turnout rate of the electorate has minimal impact on election outcomes, at least among this particular sample. However, these results must be taken with caution for four reasons. First the survey sample by necessity excludes unregistered citizens and those who are *extremely* unlikely

to vote (see Appendix I for detail). It is plausible that a simulation including these groups would yield different results, though this is hard to test given the difficulty in reaching these respondents. Second, this particular election ended with a sizable advantage for the Democrat Bob Casey. Under closer electoral conditions, turnout might play a larger role, especially if turnout levels change only for specific subgroups that are highly partisan. Third, it is plausible that my simulation method fails to capture important aspects of increased turnout. For example, could there possibly be a joint turnout/persuasion effect caused by campaigns? In the 2008 Presidential election, it seems plausible that the Obama campaign's mobilization strategy helped increase turnout as well as energize his supporters. Such a joint effect is difficult to simulate with this type of data unless I include strong assumptions. Fourth, this analysis rests on a single election in a single state in a single year. It is important to investigate more generally before reaching conclusions.

However, the overall lack of a partisan effect is apparent and found to be caused by a balance in the electorate which is not thoroughly discussed in the literature. This balance is found across multiple dimensions—partisanship, voting likelihood, and demographics. In particular, the demographic characteristics of low-voting Republicans are identified as a counter-balance to the more familiar subgroup of low-voting Democrats.

Finally, altering turnout levels for specific demographic subgroups yields smaller effects on partisan outcomes than originally anticipated, due to the size of the subgroups as a part of the full electorate, and the fact that all subgroups have some level of partisan split.

Appendix I – Survey Sampling

Because the survey respondent population voted at 75% in 2006 as compared to the Secretary of State's reported 50% turnout rate, it is important to describe the sampling methodology and certain demographic checks, in order to assert the reliability of the survey for investigating voter turnout.

The survey's intended group of respondents was registered voters, *except* voters who were a priori assumed to be extremely unlikely to vote. Specifically, the sample universe was drawn from the group of voters who either voted in the 2004 general election, or had registered in 2005 or 2006. This can be viewed as a stricter application of the idea of "deadwood," described by McDonald (2007) as the attempt to keep registration rolls accurate by purging them of people who no longer live (and vote) at their registration address.

A similar sampling strategy was performed for a statewide survey of Virginia voters in 2005 during the course of that year's gubernatorial campaign. Because that survey also included a post-election panel-back survey, it is instructive to view how turnout rates changed across populations in that instance. Table A.1 displays turnout rates for different groups, moving from the voter registration file to the group of panel-back respondents. First focus on the column labeled "2005", indicating 2005 turnout levels. Members of the full voter registration file voted at 47% in 2005. For sampling purposes, only people with phones were put into the sample universe, which raised the

	Turnout			
Dataset	**2001**	**2002**	**2004**	**2005**
Voter file (N=4,136,809)	47%	45%	76%	47%
Has phone (N=2,865,635)	53%	50%	80%	52%
Sample universe (N=200,000)	53%	50%	80%	52%
Poll respondents (N=10,009)	67%	65%	90%	69%
Panel-back respondents (N=2,856)	74%	72%	93%	81%

Table A.1 *Turnout rates for various groups, for a survey done in Virginia in 2005 with a similar sampling strategy as this survey. Each column represents turnout rates for that year's general election, as recorded on the voter registration file. For the 2005 election, the full voter registration file had a 47% turnout rate, rising to 52% for people with a phone and the sample universe. The turnout rate of poll respondents jumped to 69%, and the rate jumped again to 81% for panel-back respondents. Similar increases are seen for past years.*

---

turnout rate to 52%. The fourth row (highlighted in red) indicates that turnout rose to 69% for the poll respondents, while the fifth row indicates that turnout rose again to an astounding 81% among people who responded again to the panel-back phone interview.

These results indicate that the sampling methodology only accounting for a 5% increase in expected turnout, with most of the increase in turnout due to survey response bias. In other words, despite demographic similarity to the people in the sample universe as a whole, the mere act of completing a survey indicated a 17% increase in voting likelihood. Respondents who completed a second survey had a similar increase in expected vote.

Moving to the other columns—which indicate turnout rates for other elections based on past vote history records—notice that completing a survey increased *past* likelihood of voting by similar margins. Because this interview was conducted in May 2005, a causal interpretation—that answering a survey provides a stimulus to voting—is shown to be inappropriate.

Finally, it is important to note that both the Virginia and Pennsylvania surveys were relatively short surveys, each with an interview time of 10-15 minutes. For longer surveys, it is plausible that the increase in expected turnout among survey respondents would be even higher. Due to data availability issues, I am unable to recreate this table for the Pennsylvania file at this time. However, the magnitude of change in turnout is of similar magnitude in both surveys, suggesting that a similar process is taking place.

As an additional check on the comparability of the survey to the general population, I compare the demographic distribution of validated 2006 voters from the survey to the demographic distribution of the 2006 Pennsylvania exit poll. Table A.2 displays the results for all questions which were asked in comparable ways across the surveys. Almost all of the partisanship, vote choice, and demographic questions yield similar distributions, with two notable exceptions. First, 2006 voters from the survey used in this study (labeled "July Poll") skewed slightly older than the exit poll. Second,

		Counts		Col %		
Question	**Response**	**Exit Poll**	**July Poll**	**Exit Poll**	**July Poll**	**Diff**
Sen Vote	Bob Casey (Dem)	1,439	2,501	59%	57%	+2%
	Rick Santorum (Rep)	997	1,907	41%	43%	-2%
Gov Vote	Ed Rendell (Dem)	1,480	2,597	61%	63%	-2%
	Lynn Swann (Rep)	957	1,542	39%	37%	+2%
Party ID	Democratic	951	2,571	43%	44%	-1%
	Republican	839	2,247	38%	39%	-1%
	Independent	373	812	17%	14%	+3%
	Something else	60	175	3%	3%	-0%
Gender	Male	1,194	2,861	49%	48%	+1%
	Female	1,266	3,105	51%	52%	-1%
Age	18-24	145	145	6%	2%	+3%
	25-29	132	204	5%	3%	+2%
	30-39	339	633	14%	11%	+3%
	40-44	240	452	10%	8%	+2%
	45-49	313	515	13%	9%	+4%
	50-59	586	1,245	24%	21%	+3%
	60-64	211	534	9%	9%	-0%
	65-74	311	1,055	13%	18%	-5%
	75 or over	183	1,183	7%	20%	-12%
Race	White	2,144	5,009	88%	86%	+2%
	Black	205	552	8%	9%	-1%
	Hispanic/Latino	39	58	2%	1%	+1%
	Asian	23	34	1%	1%	+0%
	Other	28	171	1%	3%	-2%
Married	Yes	1,270	3,744	67%	63%	+4%
	No	622	2,165	33%	37%	-4%
Children in HH	Yes	753	1,542	33%	26%	+7%
	No	1,544	4,393	67%	74%	-7%
Union membership	Somehow yes	498	1,572	25%	27%	-2%
	No	1,490	4,325	75%	73%	+2%

Table A.2 *Survey comparison to the 2006 Pennsylvania exit poll. For the July poll, i.e. the survey used for this analysis, this table only uses 2006 validated voters, in order to be comparable to the respondent base for an exit poll. Most questions have a similar distribution across polls.*

---

the July poll had a slightly smaller number of respondents with children in the household. Without conducting an in-depth analysis of the weighting schemes used for both polls, it is difficult to interpret exactly why these differences may have arisen. Fortunately, the overall consistency of demographic distribution between the polls indicates that, at least for the population of 2006 voters, the July poll matches the expected distribution.

References

Brunell, Thomas L., and John DiNardo. 2004. "A Propensity Score Reweighting Approach to Estimating the Partisan Effects of Full Turnout in American Presidential Elections." *Political Analysis* 12(1):28-45.

Citrin, Jack, Eric Schickler, and John Sides. 2003. "What If Everyone Voted? Simulating the Impact of Increased Turnout in Senate Elections." *American Journal of Political Science* 47(1):75-90.

DeNardo, James. 1980. "Turnout and the Vote: The Joke's on the Democrats." *American Political Science Review* 74(2):406-420.

Erikson, Robert S. 1995. "State Turnout and Presidential Voting." *American Politics Quarterly* 23(4):387-396.

Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

Green, Donald P. and Alan S. Gerber. 2006. "Can Registration Based Sampling Improve the Accuracy of Midterm Election Forecasts?" *Public Opinion Quarterly* 70(2):197–223.

Leighley, Jan E., and Jonathan Nagler. 2007. "Who Votes Now? And Does It Matter?" Presented at the 2007 Annual Meeting of the Midwest Political Science Association.

Malchow, Hal. *The New Political Targeting.* Predicted Lists, LLC, 2008.

Martinez, Michael D., and Jeff Gill. 2005. "The Effects of Turnout on Partisan Outcomes in U.S. Presidential Elections 1960-2000." *The Journal of Politics* 67(4):1248-1274.

McDonald, Michael P. 2007. "The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics." *Public Opinion Quarterly* 71(4):588-602.

Nagel, Jack H., and John E. McNulty. 1996. "Partisan Effects of Voter Turnout in Senatorial and Gubernatorial Elections." *American Political Science Review* 90(4):780-793.

Radcliff, David. 1994. "Turnout and the Democratic Vote." *American Politics Quarterly* 22(3):259-276.

Wolfinger, Raymond E., and Steven J. Rosenstone. *Who Votes?* Yale University Press. 1978.