Modeling Voter Turnout under the Habitual Voter Theorem

Richard Wang
Washington University in St. Louis, rtwang@wustl.edu

5 June 2017

Abstract **- This paper uses logistic regression to model voting behavior under the habitual voter theorem. Data is provided by the North Carolina State Board of Elections. The model finds that the total number of previous votes an individual cast significantly influences the probability of that individual turning out to vote in the 2016 presidential election. Specifically, each of the increases in the individual's previous total vote from zero to one, one to two, two to three, three to five, and five above increases the likelihood of that individual voting in the 2016 election by 1.868 times ($p < 0.001$). The results of this paper do not necessarily prove that voting is habit-forming; rather, they suggest that those who have participated in previous elections are more likely to turn out in future ones (under the existing framework of the habitual voter theorem). This paper also provides an interactive model that measures the probability of an individual voter turning out in the 2016 presidential election given a particular set of characteristics. A description of the probability model can be found in Appendix B.**

"Laws are never as effective as habits"-Adlai Stevenson

Introduction

The United States suffers from relatively low turnout rates: voter turnout rates for presidential and midterm elections consistently hover around 60% and 40% respectively [1]. Down's (1957) rational choice theory suggests that turnout rates should be even lower, arguing that since the chances of casting the pivotal vote is essentially nil, the benefits of voting never outweigh the costs [2]. Recently, however, scholars are beginning to recognize the social capital of voting, theorizing that the act of voting may provide other, social benefits such as civic fulfillment or the avoidance of social sanction [3, 4]. Individuals who perceive such social benefits to outweigh the costs of voting (i.e. driving to the local polling center and waiting in line) are consequently more likely to turnout to vote. Furthermore, recent literature suggests that voting may be habit-forming, arguing that the (real and perhaps perceived) costs of voting decrease as a result of previous experiences. Under this framework, we would expect that the chances of an individual participating in an election increases in relation to the number of times that individual has voted in the past. This paper introduces a logistic model of voter turnout building on such a theory.

The Habitual Voter Theorem

The basic premise of the habitual voter theorem suggests that voting is habit-forming and conceptually stems from the "familiarity heuristic" which was first tested by Amos Tversky and Daniel Kahneman in their 1973 paper "Judgement under Uncertainty: Heuristics and Biases." The familiarity heuristic occurs "when the familiar is favored over novel places, people, things" [5]. Although Tversky and Kahneman were able to examine the phenomenon during their study, their explanation of it was rather lacking. However, the familiarity heuristic is conceptually similar to that of the mere exposure effect, which was both defined and adequately explained by Robert Zajonc in his 1968 paper "Attitudinal Effects of Mere Exposure." Summarizing Rajonc's work in a concise manner, Kahneman wrote the following:

> Zajonc argued that the effect of repetition on liking is a profoundly important biological fact... To survive in a frequently dangerous world, an organism should react cautiously to a novel stimulus, with withdrawal and fear… However, it is also adaptive for the initial caution to fade if the stimulus is actually safe. The mere exposure effect occurs… because the repeated exposure of a stimulus is followed by nothing bad. Such a stimulus will eventually become a safety signal, and safety is good [6].

Although such an abstract concept hardly seems relevant in the context of voting, its influence should not be dismissed. Individuals, especially those who are risk averse and who have never voted before, may choose not to participate in the electoral process simply because they naively perceive the costs of voting to be high. Indeed, the image of the struggling voter waiting in line for hours in the rain is likely to capture the minds of many, when in reality the national average wait time is 14 minutes (mail-in ballots require even less time) [7]. Scholars Kevin Denny and Orla Doyle offer further insight to the influence of familiarity in their article "Does Voting History Matter? Analyzing Persistence in Turnout," arguing that the initial costs of voting are initially high because one has to "find the polling station, learn how to cast a vote and differentiate between political parties." Consequently, once an individual has leaped the initial hurdle, his or her costs will be lower for subsequent elections, thus increasing the likelihood that s/he will vote in future elections [8]. Denny and Doyle continue to suggest that "participating in an election… enhances the voters' interest in politics and increases their sense of civic duty, all of which strengthen the positive connotation of voting" [8].

Moving on from theory, existing literature appears to affirm the habitual voter theorem and its application in explaining actual voter behavior. Using 1972 and 1992 American National Election Study Panel Surveys (ANES) data, Donald P. Green and Ron Shachar (2000) were able to develop models evaluating the habitual voter theorem while controlling for a multitude of factors, such as closeness of election, evaluation of the candidates, political interest, education, and other demographic characteristics. Green and Shachar concluded that the effect of past voting on an individual's decision to vote in a future election is "quite large" and provided the following illustration:

> …consider hypothetical voters who have a 50% probability of going to the polls on election day. Using the median probit coefficient of .93 for purposes of illustration, we calculate that if these voters vote in a given election, their probability of voting in the next climbs from 50% to 82% [9].

Consequently, the habitual voter theorem appears to hold merit. Further research by Green, Sachar, and Alan S. Gerber found similar results. In their article "Voting May Be Habit-Forming: Evidence from a Randomized Field Experiment," Green, Sachar, and Gerber develop a model that "allows for both unobserved heterogeneity among individuals and the potential force of habit" (542). Since voting in 1998 is "potentially correlated with unmodeled causes of voting in 1999," Gerber et al. used a two-stage least-squares regression involving the treatment effects of direct mail and personal canvassing on turnout (determined by their 1998 field study) to isolate the effect of habit. From their regression, they found that "voting in 1998 raised the probability of voting in 1999 by 46.7 percentage points" (547). Such effects of habit surpassed even those of other (traditional) demographic variables, such as education and age.

Hypotheses

We hope to contribute to this field of study by analyzing voter files provided by the North Carolina State Board of Elections. Although our study employs primitive statistical analytics, it is still able to assess the validity of the habitual voter theorem to a degree; furthermore, our study aims to consider the implications of participating in different types of elections (presidential versus midterm) rather than elections in general. Although our results will not be without faults, they will at least provide a novel perspective to the field and the data used.

As demonstrated by both theory and previous literature, we would expect the probability turnout to increase as past vote count increases. Thus:

1) Individuals with higher vote counts between the years 2007 and 2015 are more likely to vote in the 2016 presidential election than those with lower vote counts.

We would also expect individuals who frequently participate in midterm elections to be even more likely to turnout in future elections. Midterm elections often lack the extensive media coverage and "wow factor" that presidential elections possess (see Campbell 1987). Consequently, the cost of turning out to vote in midterm elections is relatively higher as candidates (and their positions) are usually less well known to the populace than presidential candidates in respect to presidential elections. We suspect that individuals voting in midterm elections perceive the costs of voting to be lower and/or the (social) payoff of voting to be higher than individuals who do not participate in midterm elections. Thus:

2) Individuals with a higher percentage of votes cast in off-presidential midterm years within their total vote count are more likely to vote in the 2016 presidential election than those with a lower percentage.

Lastly, we suspect that a vote in the preceding presidential election will act as a strong indicator of whether an individual chooses to vote in the 2016 general election (simply because of the recency and thus familiarity with the phenomenon). Thus:

3) Individuals who voted in the 2012 presidential election are more likely to vote in the 2016 presidential election than those who did not.

Data and Methods

As mentioned before, longitudinal voter data is provided by the North Carolina State Board of Elections.
(link: http://dl.ncsbe.gov/index.html?prefix=data/; last updated: 05/18/2017).

The files provided include information on each voters' vote history along with their demographic information such as race, party affiliation, gender, ethnicity, and age. The vote history includes presidential, primary, municipal, and other* elections between the years 2007 and 2016.[1] Each voter also had a unique "ncid" which we used to match voting history to the respective voter. From this data, we create four variables:

1. **Total Vote**: The total number of times a voter had voted before the 2016 presidential election (which includes the 2016 primaries).
2. **Midterm Percentage**: Midterm Voter/Total Vote.

   Note: The Midterm Vote is the total number of times a voter had voted in an off-presidential year election (i.e. years [2007,2016] excluding 2008, 2012, and 2016).
3. **Preceding Presidential**: A boolean variable with 1 indicating that the voter had voted in the 2012 presidential election and a 0 otherwise.
4. **2016 General**: A boolean variable with a 1 indicating that the voter had voted in the 2016 presidential election and a 0 otherwise.

Further manipulation of the dataset was needed however. We removed any voter under the age of 27, thus excluding those who were under the age of 18 in the year 2007. Since these voters were ineligible to participate in one or more of the elections prior to the 2016 presidential elections (and thus did not receive "treatment"), their participation in the study would have produced skewed results. Voters also had a status designating them as "ACTIVE," "INACTIVE," or "REMOVED." We removed any voters with the status "REMOVED" since such status usually indicated that the voter was either deceased or no longer eligible to vote (i.e. felons). Removing these voters was necessary for the same concern as noted above. Any vote histories which could not be matched to a voter with valid demographic inputs (i.e. marked with "NA") were also removed. After the removals, we were left with a pool of 4,951,648 voters from an initial pool of 6,586,179.

Using the voter data, we then created the following control variables:

1. **Race**: Asian, African-American, Indian-American, Two or More, Other, Undesignated, and White.
2. **Party**: Democrat, Libertarian, Republican, and Unaffiliated.
3. **Gender**: Female, Male, Undesignated.
4. **Age Group**: 27-34, 35-44, 45-54, 55-64, 65+.

   The age groups were determined by suggested survey classifications (link: http://www.pgagroup.com/standardized-survey-classifications.html).
5. **Ethnicity:** Hispanic, not-Hispanic, and Unknown.

We were also able to pull census data using the zip code information provided within the voter files (census data was provided by the following site: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml). The census data provided the following derived information at the zip code level:

6. **Pct_Under20**: The percentage of population younger than 20 years of age.
7. **Pct_Under40**: The percentage of population between the age of 20 and 39, inclusive.
8. **Pct_Under60**: The percentage of population between the age of 40 and 59, inclusive.
9. **Pct_Above60**: The percentage of population 60 years of age or older.
10. **Pct_Hisp**: The percentage of Hispanic population.
11. **Pct_White**: The percentage of White population.
12. **Pct_Black**: The percentage of Black population.
13. **Pct_Asian**: The percentage of Asian population.
14. **Pct_Other**: The percentage of another identified ethnicity population.
15. **Pct_Renter**: The percentage of individuals renting their houses.
16. **Pct_Owner**: The percentage of individuals owning their houses.
17. **House_Incom_Below55k**: A boolean variable: 1 if the average household income for the zip code is below 55k, 0 otherwise.
18. **Avg_House_Size**: The average household size.
19. **Pct_Blue_Collar**: The percentage of blue collar workers (population = civilian employed population who are 16 years and older).
20. **Pct_White_Collar**: The percentage of white collar workers (population = civilian employed population who are 16 years and older).

[1] An election that was not marked as presidential, primary, nor municipal was categorized as "other."

The following census variables were also available but were ultimately excluded due to issues with collinearity (see **Appendix A**):

21. **Avg_House_Inc:** The average household income (highly correlated with Pct_White_Collar).
22. **Avg_Inc**: The average income per capita (highly correlated with Avg_House_Inc).
23. **Avg_House_Value**: The average house value (highly correlated with Avg_House_Inc).
24. **Pct_BS_Educ**: The percentage of individuals over the age of 25 with a B.S. degree or higher (highly correlated with Pct_White_Collar).
25. **Pct_College_Educ**: The percentage of individuals over the age of 25 with some college education.
26. **Pct_HS_Educ**: The percentage of individuals over the age of 25 with high school or under education.

The model used in this study is the logistic regression, using Total Vote, Midterm Percentage, and Preceding Presidential as our predictor variables and 2016 General as our response variable. The model also controlled for variables 1-20, as mentioned above.

$$\text{Log}(p/1\text{-}p) = b_0 + b_1(\text{Total Vote}) + b_2(\text{Midterm Percentage}) + b_3(\text{Preceding Presidential}) + b_4(\text{Race})... + b_{23}(\text{Pct\_White\_Collar}) + e$$

Results

With our model, we attempt to answer three questions (keeping our earlier hypotheses in mind): Firstly, which factors influence the 2016 election vote? Secondly, among those factors, which have significant influence? Lastly, which combination of factors drives the highest outcome?

Total Vote appears to play a role in turnout:

FIGURE 1 illustrates the 2016 presidential election turnout rates for each group of voters (determined by the Total Vote value). For example, 54.2% of individuals who had previously voted only once turned out to vote in the 2016 presidential election. We see that as the Total Vote increases, the turnout rate increases as well: turnout rates for individuals who had previously voted more than 16 times exceed 99%, a drastic jump from the earlier 54%. Such patterns exist across all subgroupings of Race, Party, Gender, and Age Group against Total Vote, although there appears to be an outlier for the Libertarian group against Total Vote (see FIGURE 2 - 5 under **Figures and Tables** for more information; turnout tables were not produced for any of the census variables as they were at the zip code – rather than individual – level). We purposefully excluded

FIGURE 1

Total Vote	Turnout %
1	54.2%
2	69.8%
3	80.0%
4	86.5%
5	91.0%
6	93.8%
7	95.3%
8	96.1%
9	96.8%
10	97.3%
11	97.6%
12	97.9%
13	98.2%
14	98.5%
15	98.6%
16	99.0%
17	99.1%
18	99.1%
19	99.4%
20	99.4%
21	99.6%
22	99.6%
23	99.8%
24+	100.0%

the group of voters with Total Vote = 0 since individuals in this group have only voted in the 2016 presidential elections (only individuals who have voted at least once between the years 2007 and 2016 are listed in North Carolina's voter files – nonvoters are excluded).

The first run of our logistic regression produced statistically significant coefficients for all of our predictor variables, all in the correct direction. However, we noticed that, as illustrated by FIGURE 1, the effects per each increase of Total Vote were not static (that is, the relationship was not linear, but rather logarithmic).

Consequently, we devised buckets for both our Total Vote variable and Midterm Percentage variable:

1. **Total Vote**: The buckets for the revised Total Vote variable are the following: 1, 2, 3, (4 or 5), 5+.
2. **Midterm_Pct_Over25**: A boolean variable, 1 if the Midterm Percentage is greater than 25%, 0 otherwise.

The results for the rerun of our logistic regression are shown below:

TABLE 1 & TABLE 2 GO HERE

1) Which factors influence the 2016 election vote?

Quickly summarizing some of the results of our control variables (see the coefficients listed in TABLE 1), we see that individuals of every race are less likely to vote than White voters except for Asian voters (who are more likely to turnout) and racially-undesignated voters (who turnout at essentially the same rate compared to Whites); voters registered as Libertarian, Republican, or

Unaffiliated were all more likely to turnout than those registered as Democrat; male voters were more likely to turnout than female and undesignated voters; voters in the age groups 34-44, 45-54, and 55-64 were more likely to turnout than voters in the age group 27-34 while voters in the age group 65+ were less likely; Hispanic voters were more likely to vote than non-Hispanic voters, while ethnically-unknown were less likely; and so forth.

As observed in TABLE 1, the coefficients for our predictor variables (Total Vote, Midterm Percentage, and Preceding Presidential) were all statistically significant at the $p < 0.001$ level, and in the correct (positive) direction. For Total Vote, an increase from 0 to 1, 1 to 2, 2 to 3, 3 to (4 or 5), and 5+ all each increased the likelihood of that voter turning out in the 2016 presidential election by 1.868 times. For Preceding Presidential, we see that if a voter participates in the 2012 presidential election, the chances of his or her voting in the 2016 presidential election increases 1.494 times. Lastly, if the Midterm Percentage (Midterm Vote / Total Vote) is greater than 25%, then that voter is 1.255 times more likely to turnout in the 2016 presidential election. All of these interpretations involve the odds ratios (see TABLE 2).

The cumulative effect of all three variables, calculating as Total Vote from 0 to 4, Preceding Presidential from 0 to 1, and Midterm Percentage from 0 to 1 is over 22 times (see calculation in TABLE 3), which indicates that the odds for an individual who has total vote of 4 (with 1 of them from preceding presidential election and 1 of them from midterm) voting in the 2016 presidential election is significantly higher.

2) Which of the aforementioned factors have the greatest influence on the 2016 election vote?

The odds ratios (OR) indicate the influence of the variables on turnout in the 2016 presidential election: the greater the difference between the OR and 1, the greater the influence that variable possesses (a value of 1 would indicate no effect on outcome). The accuracy of the OR is indicated by the confidence interval (CI): a large CI indicates a low level of precision of the OR, whereas a small CI indicates a higher precision of the OR. We create the following measurement to rank the variables by their influence: ABS(OR-1).[2] Total Vote, Preceding Presidential, and Midterm_Pct_Over25 rank as the 1st, 3rd, and 7th most influential variables (see TABLE 4). Using the statistical R package "caret" (developed by Max Kuhn, link: https://topepo.github.io/caret/index.html) which ranks variables by the importance of their contribution to regression models, we are able to affirm that Total Vote, Preceding Presidential, and Midterm_Pct_Over25 rank 1st, 2nd, and 5th (see TABLE 5).

3) Which combination of factors drives the highest outcome?

Using the logistic regression output, we build an interactive model that provides the probability of an individual voter with particular characteristics turning out to vote in the 2016 presidential election (see **Appendix B** for the model details). We then create a hypothetical voter who is male, white, Republican, between the ages of 35 and 44, non-Hispanic, and lives with a household income above 55k; for practical purposes of this hypothetical, we have used state averages as inputs for the remaining census variables. The hypothetical voter also has Total Vote, Preceding Presidential, and Midterm_Pct_Over25 all equal to zero; consequently, his probability of turning out in the 2016 presidential election is 36.2%.

Changing the hypothetical voter's Total Vote from 0 to 1 increases the voter's probability of turning out from 36.2% to 51.5%. Further increasing his Total Vote to 5 increases his probability of turning out to 92.8%.

Resetting the voter's Total Vote back to 0 and changing his Preceding Presidential variable from a 0 to a 1, the voter's probability of turning out increases from 36.2% to 45.9%.

Resetting the voter's Preceding Presidential variable back to 0 and changing his Midterm_Pct_Over25 from a 0 to a 1, the voter's probability of turning out increases from 36.2% to 41.7%

Attempting to maximize the likelihood of turnout, we then increase his Total Vote to 5 and we change both Preceding Presidential and Midterm_Pct_Over25 to one to calculate the scenario in which a voter is most likely to vote – with a probability of turning out at 96% (a 59.8% increase in probability of voting).

Limitations

There are several limitations to our study. Firstly, while our model uses the habitual voter theorem to model voter behavior, it does not (nor does it attempt to) measure the exact effects of habit on voting. The model only uses past voting behavior as a predictor of future voting behavior. As previously mentioned, Gerber et al. have argued that a simple regression on lagged voting is insufficient to determine the influence of habit on voting due to the possibility of external variables effecting both past and future turnout. Consequently, readers should not interpret our model as if it were proving the concept of habitual voting itself.

Secondly, besides the Preceding Presidential variable, the model did not account for a temporal relationship between

[2] ABS() is the absolute function.

past votes and participation in the 2016 election. For example, votes occurring in more recent years may have held greater influence on an individual's decision to vote in 2016 than votes occurring in the more distant past. However, the regression model treated these votes similarly – as a cumulative count (Total Vote).

Other limitations exist within the North Carolina voter files. The data does not include crucial information such as voters' education and income levels. We were able to mitigate this issue by utilizing census data offered at the zip code level; an improvement would be to rerun the model with individual level data, along with other information such as residential mobility, exposure to campaign ads, political interest, weather, etc. (i.e. survey data). Furthermore, the data does not include the nonvoter population – individuals who have never voted in any election (including the 2016 presidential election). Consequently, we had to remove the population of voters with Total Vote = 0 from our analysis. Although the issue is mitigated by the large population size of the remaining voter pool, it remains as a limitation nonetheless.

Lastly, North Carolina's status as a swing state is also likely to introduce concerns to our model. Since the election is considered to be more competitive, it is more likely to receive a greater amount of media coverage. Furthermore, individuals may be more likely to turnout merely because they perceive their votes to count more. While this does not pose a limitation to our model, since all subjects observed (North Carolina voters) are exposed to the influences, it does pose as a challenge to any attempt to generalize the model's results to extra-state elections.

Conclusion

Individuals who voted in the past were significantly more likely to vote in the 2016 presidential election. The starkest example of this is observed by our previous approximation of a "typical" voter, whose probability of turning out in the 2016 presidential election increased by 56.6% when his previous total vote count increased from 0 to 5. Other factors such as whether the individual voted in the 2012 presidential election and the midterm election percentage also played statistically significant roles in explaining voter turnout. Furthermore, a voter only needs to have previously voted two or three times to be significantly more likely to vote in the 2016 election. The probability of voting in the 2016 presidential elections for individuals who have previously voted in two or three elections are 66.48% and 78.75% respectively – a drastic increase from the initial 36.2%.

As previously mentioned, the model takes the habitual voter theorem for granted – the model does not measure the effects of habit on turnout but the influence of previous voting behavior on future voting behavior. It is entirely possible that these voters are consistently voting because they are consistently targeted by campaign advertisements. However, we follow previous literature and assume that voting is truly habit-forming. Furthermore, this study makes no attempt to recommend particular strategies for voter engagement campaigns to increase voter turnout – an undoubtedly difficult task best left to its respective experts. The results of this study simply suggest that voters only need to be persuaded into voting in two or three elections until they become consistent voters.

Acknowledgments

Special thanks to Ryden Butler for providing analytical and content review and to Lilly Wang for providing statistical guidance.

References

[1] FairVote.org. "Voter Turnout." FairVote. FairVote, n.d. Web. 4 June 2017.

[2] Farber, Henry S. Rational choice and voter turnout: Evidence from union representation elections. No. w16160. National Bureau of Economic Research, 2010.

[3] Atkinson, Matthew D., and Anthony Fowler. "The Effect of Social Capital on Voter Turnout: Evidence from Saint's Day Fiestas in Mexico." Conference of the Midwest Political Science Association. 2011.

[4] Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. "Social pressure and voter turnout: Evidence from a large-scale field experiment." American political Science review 102.1 (2008): 33-48.

[5] Tversky, Amos, and Daniel Kahneman. "Judgment under uncertainty: Heuristics and biases." Utility, probability, and human decision making. Springer Netherlands, 1975. 141-162.

[6] Zajonc, Robert B. "Attitudinal effects of mere exposure." Journal of personality and social psychology 9.2p2 (1968): 1.

[7] Bump, Philip. "Planning to Vote? Here's How Long You Could Wait." The Washington Post. WP Company, 03 Nov. 2014. Web. 05 May 2017.

[8] Denny, Kevin, and Orla Doyle. "Does voting history matter? Analysing persistence in turnout." American Journal of Political Science 53.1 (2009): 17-35.

[9] Green, Donald P., and Ron Shachar. "Habit formation and political behaviour: Evidence of consuetude in voter turnout." British Journal of Political Science 30.04 (2000): 561-573.

[10] Gerber, Alan S., Donald P. Green, and Ron Shachar. "Voting may be habit-forming: evidence from a randomized field experiment." American Journal of Political Science 47.3 (2003): 540-550.

[11] Campbell, James E. "The revised theory of surge and decline." American Journal of Political Science (1987): 965-979.

Figures and Tables

FIGURE 1

Total Vote	Turnout %
1	54.2%
2	69.8%
3	80.0%
4	86.5%
5	91.0%
6	93.8%
7	95.3%
8	96.1%
9	96.8%
10	97.3%
11	97.6%
12	97.9%
13	98.2%
14	98.5%
15	98.6%
16	99.0%
17	99.1%
18	99.1%
19	99.4%
20	99.4%
21	99.6%
22	99.6%
23	99.8%
24+	100.0%

FIGURE 2

	Race						
Total Vote	Asian	Black	American Indian	Two or More	Other	Undesignated	White
1	61.4%	45.6%	38.0%	47.5%	57.0%	57.1%	56.8%
2	75.4%	64.0%	55.4%	64.0%	72.0%	73.3%	71.7%
3	84.3%	76.8%	68.9%	76.2%	80.5%	82.4%	81.1%
4	88.0%	84.7%	76.5%	82.4%	86.7%	87.6%	87.1%
5	91.7%	90.4%	82.9%	87.5%	90.5%	91.4%	91.3%
6	94.2%	93.5%	87.4%	91.5%	93.5%	93.8%	94.0%
7	95.5%	95.3%	89.9%	93.7%	94.8%	94.7%	95.3%
8	95.1%	96.3%	93.1%	93.9%	95.3%	95.7%	96.1%
9	98.6%	96.7%	95.6%	94.8%	96.0%	95.8%	96.9%
10	96.2%	97.3%	95.6%	97.6%	96.6%	97.5%	97.3%
11	95.2%	97.6%	95.5%	96.1%	97.2%	96.8%	97.6%
12	97.0%	97.9%	97.3%	97.4%	97.9%	97.1%	97.9%
13	96.5%	98.1%	94.0%	98.0%	98.8%	96.1%	98.2%
14	100.0%	98.5%	94.7%	98.9%	98.6%	99.0%	98.5%
15	98.0%	98.4%	97.0%	100.0%	99.5%	99.7%	98.7%
16	98.0%	99.2%	95.4%	88.4%	98.8%	98.9%	99.0%
17	100.0%	99.0%	97.6%	100.0%	96.7%	98.7%	99.1%
18	100.0%	99.1%	95.2%	100.0%	100.0%	98.2%	99.2%
19	100.0%	99.4%	100.0%	100.0%	97.7%	98.5%	99.5%
20	100.0%	99.5%	100.0%	100.0%	97.1%	100.0%	99.4%
21	100.0%	99.6%	100.0%	100.0%	100.0%	100.0%	99.7%
22	100.0%	99.8%		100.0%	100.0%	100.0%	99.6%
23	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.6%
24+		100.0%			100.0%	100.0%	100.0%

Figures and Tables (Continued)

FIGURE 3

Total Vote	Party			
	Democrat	Libertarian	Republican	Unspecified
1	49.0%	55.9%	58.6%	56.2%
2	65.4%	69.6%	73.9%	71.6%
3	76.9%	80.4%	82.9%	81.3%
4	84.1%	85.7%	88.7%	87.4%
5	89.4%	90.4%	92.7%	91.2%
6	92.7%	92.6%	94.9%	94.0%
7	94.7%	91.9%	95.8%	95.6%
8	95.7%	93.1%	96.5%	96.4%
9	96.4%	94.1%	97.1%	97.2%
10	96.9%	96.8%	97.6%	97.5%
11	97.1%	95.5%	97.9%	97.9%
12	97.7%	98.4%	98.1%	98.1%
13	97.9%	100.0%	98.4%	98.4%
14	98.4%	100.0%	98.6%	98.7%
15	98.5%	100.0%	98.7%	98.9%
16	99.0%	100.0%	99.0%	99.3%
17	98.9%	100.0%	99.3%	99.1%
18	99.1%	100.0%	98.9%	99.5%
19	99.4%	66.7%	99.4%	99.6%
20	99.4%	100.0%	99.3%	99.5%
21	99.6%	100.0%	99.8%	99.5%
22	99.6%		99.3%	100.0%
23	99.8%		100.0%	99.3%
24+	100.0%		100.0%	100.0%

FIGURE 4

Total Vote	Gender		
	Female	Male	Unspecified
1	55.1%	52.9%	57.8%
2	70.1%	69.3%	73.4%
3	80.0%	80.0%	82.2%
4	86.2%	86.8%	87.8%
5	90.8%	91.3%	91.0%
6	93.6%	94.0%	93.6%
7	95.1%	95.5%	95.2%
8	95.9%	96.4%	95.3%
9	96.7%	97.1%	96.5%
10	97.1%	97.5%	97.8%
11	97.4%	97.8%	97.1%
12	97.7%	98.1%	97.4%
13	98.1%	98.3%	97.9%
14	98.3%	98.7%	98.8%
15	98.5%	98.7%	99.3%
16	99.0%	99.1%	98.9%
17	99.0%	99.2%	98.5%
18	99.2%	99.1%	97.2%
19	99.4%	99.4%	100.0%
20	99.3%	99.5%	100.0%
21	99.7%	99.6%	100.0%
22	99.7%	99.5%	100.0%
23	99.9%	99.6%	100.0%
24+	100.0%	100.0%	

FIGURE 5

	Age Group				
Total Vote	25-34	35-44	45-54	55-64	65+
1	49.4%	54.4%	56.7%	58.7%	53.0%
2	65.2%	70.8%	73.0%	73.6%	65.0%
3	77.0%	81.5%	83.1%	83.0%	74.0%
4	84.7%	87.9%	89.3%	89.2%	80.9%
5	89.7%	92.2%	93.2%	93.1%	86.9%
6	93.0%	94.9%	95.7%	95.4%	90.7%
7	94.3%	96.3%	96.7%	96.8%	92.9%
8	95.8%	96.9%	97.5%	97.5%	94.3%
9	96.3%	97.4%	98.0%	98.1%	95.5%
10	96.6%	97.9%	98.2%	98.3%	96.3%
11	97.6%	98.2%	98.4%	98.5%	96.8%
12	97.6%	98.6%	98.7%	98.8%	97.2%
13	97.1%	98.4%	99.1%	98.8%	97.6%
14	97.9%	99.1%	99.2%	99.2%	98.0%
15	97.7%	99.3%	99.1%	99.3%	98.2%
16	97.4%	99.2%	99.6%	99.3%	98.9%
17	99.1%	99.1%	99.4%	99.4%	98.9%
18	98.9%	99.0%	99.3%	99.4%	99.0%
19	100.0%	99.4%	99.8%	99.6%	99.3%
20	100.0%	100.0%	99.5%	99.6%	99.3%
21	100.0%	100.0%	100.0%	99.9%	99.5%
22	100.0%	100.0%	99.5%	100.0%	99.5%
23	100.0%	100.0%	100.0%	99.6%	99.8%
24+		100.0%	100.0%	100.0%	100.0%

Tables and Figures (Continued)

TABLE 1

Logistic regression output

Deviance Residuals:	**Min**: -2.8114	**1Q**: 0.2718	**Median**: 0.3566	**3Q**: 0.5415	**Max**: 1.8672
Dependent Variable					
2016 General					
Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.668638	0.18274	-20.076	< 2e-16	***
Predictor Variables:					
Total Vote	0.624869	0.002197	284.417	< 2e-16	***
Preceding Presidential	0.401873	0.005766	69.694	< 2e-16	***
Midterm_Pct_Over25	0.227919	0.005611	40.623	< 2e-16	***
Race (with 'White' as reference):					
Asian	0.126007	0.021076	5.979	2.25E-09	***
African-American	-0.147048	0.006584	-22.333	< 2e-16	***
Indian-American	-0.296352	0.02645	-11.204	< 2e-16	***
Two or More	-0.295859	0.029507	-10.027	< 2e-16	***
Other	-0.043429	0.017136	-2.534	0.011265	*
Undesignated	0.026339	0.01577	1.67	0.094867	.
Party (with 'Democrat' as reference):					
Libertarian	0.104778	0.034481	3.039	0.002376	**
Republican	0.223923	0.006114	36.625	< 2e-16	***
Unaffiliated	0.150605	0.00592	25.44	< 2e-16	***
Gender (with 'Male' as reference):					
Female	0.046913	0.004479	10.474	< 2e-16	***
Undesignated	0.181087	0.020086	9.015	< 2e-16	***
Age Group (with '27-34' as reference):					
34-44	0.206618	0.007199	28.699	< 2e-16	***
45-54	0.332455	0.0073	45.541	< 2e-16	***
55-64	0.408585	0.007646	53.437	< 2e-16	***
65+	-0.06172	0.007367	-8.378	< 2e-16	***
Ethnicity (with 'not-Hispanic' as reference):					
Hispanic	0.203101	0.018094	11.224	< 2e-16	***
Unknown	-0.061083	0.005708	-10.702	< 2e-16	***
Zip Code Age (with 'Pct_Under20' as complement):					
Pct_Under40	0.006477	0.001650	3.925	8.67E-05	***
Pct_Under60	0.020495	0.001791	11.443	< 2e-16	***
Pct_Above60	0.005355	0.001513	3.539	0.000402	***
Zip Code Ethnicity (with 'Pct_Other' as complement):					
Pct_Hisp	0.009564	0.000693	13.806	< 2e-16	***
Pct_White	0.008784	0.000464	18.919	< 2e-16	***
Pct_Black	0.008382	0.000486	17.241	< 2e-16	***
Pct_Asian	0.011895	0.001121	10.614	< 2e-16	***
Zip Code Housing (with 'Pct_Renter' as complement):					
Pct_Owner	0.004546	0.000450	10.098	< 2e-16	***

Tables and Figures (Continued)

TABLE 1 ***(Continued)***

Logistic regression output

Zip Code Household and Income:					
Avg_House_Size	0.174605	0.030613	5.704	1.17E-08	***
House_Income_Below55k	-0.077051	0.006564	-11.738	< 2e-16	***
Zip Code Occupation (with 'Pct_Blue_Collar' as complement):					
Pct_White_Collar	0.003645	0.000359	10.153	< 2e-16	***

Signif. codes: ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 1637072 on 1809970 degrees of freedom					
Residual deviance: 1340913 on 1809939 degrees of freedom					
AIC: 1340977					

Tables and Figures (Continued)

TABLE 2

Logistic regression output with odds ratio*

	Confidence Interval		
	Odds Ratio	lower	upper
Predictor Variables			
Total Vote	1.868001	1.859975	1.876062
Preceding Presidential	1.494621	1.477824	1.511609
Midterm_Pct_Over25	1.255984	1.242247	1.269872
Race (with 'White' as reference):			
Asian	1.134290	1.088387	1.182128
Black	0.863253	0.852184	0.874465
Indian-American	0.743526	0.705962	0.783088
Two or More	0.743892	0.702091	0.788182
Other	0.957501	0.925875	0.990206
Party (with 'Democrat' as reference):			
Libertarian	1.110465	1.037895	1.188108
Republican	1.250975	1.236074	1.266056
Unaffiliated	1.162537	1.149126	1.176105
Gender (with 'Male' as reference):			
Female	1.048031	1.038870	1.057271
Undesignated	1.198519	1.152251	1.246645
Age Group (with '27-34' as reference):			
35-44	1.229513	1.212285	1.246985
45-54	1.394387	1.374578	1.414481
55-64	1.504687	1.482305	1.527406
65+	0.940146	0.926668	0.953820
Ethnicity (with 'not-Hispanic' as reference):			
Hispanic	1.225196	1.182505	1.269427
Unknown	0.940746	0.930280	0.951329
Zip Code Age (with 'Pct_Under20' as complement):			
Pct_Under40	1.006498	1.003248	1.009759
Pct_Under60	1.020707	1.017130	1.024296
Pct_Above60	1.005369	1.002392	1.008355
Zip Code Ethnicity (with 'Pct_Other' as complement):			
Pct_Hisp	1.009610	1.008240	1.010982
Pct_White	1.008822	1.007905	1.009741
Pct_Black	1.008417	1.007457	1.009379
Pct_Asian	1.011966	1.009746	1.014192

Tables and Figures (Continued)

TABLE 2 ***(Continued)***

Logistic regression output with odds ratio

Zip Code Housing (with 'Pct_Renter' as complement):			
Pct_Owner	1.004557	1.003671	1.005443
Zip Code Household and Income:			
Avg_House_Size	1.190776	1.121429	1.264411
House_Income_Below55k	0.925843	0.914008	0.937831
Zip Code Occupation (with 'Pct_Blue_Collar' as complement):			
Pct_White_Collar	1.003652	1.002946	1.004359

***** The percentage variables are in percent units. The odds ratios for them apply to each increase in percent units.

TABLE 3

Variance/Covariance of Total Vote, Preceding Presidential, and Midterm Percentage

	Coefficient	Total Vote	Preceding Presidential	Midterm_Pct_over25
Total Vote	0.624869	0.00000483	-0.00000634	-0.00000697
Preceding Presidential	0.401873	-0.00000634	0.00003325	0.00000746
Midterm_Pct_over25	0.227919	-0.00000697	0.00000746	0.00003148

The OR of 2016 presidential election turnout comparing Total Vote of 4, Preceding Presidential of 1 and Midterm Percentage of 1 to Total Vote of 0, Preceding Presidential of 0 and Midterm Percentage of 0 is calculated as follows:

OR = $e^{(4*0.624869+0.401873+0.227919)} = 22.86$

$$SE = \sqrt{\text{var}(\beta_1) + \text{var}(\beta_2) + \text{var}(\beta_3) + 2\,\text{cov}(\beta_1 \quad \beta_2) + 2\,\text{cov}(\beta_1 \quad \beta_3) + 2\,\text{cov}(\beta_2 \quad \beta_3)}$$

$$= \sqrt{0.00000483 + 0.00003325 + 0.00003148 - 2 * 0.00000634) - 2 * 0.00000697 + 2 * 0.00000746}$$

$$= 0.007606$$

CI = $e^{(4*0.624869+0.401873+0.227919)\pm1.96*0.007606} = 22.52 – 23.20$

Tables and Figures (Continued)

TABLE 4

Variable influence ranked by difference between odds ratio and 1

Variable	Odds Ratio	ABS(1-OR)	Variable Associated with
Total Vote	1.868000972	0.868000972	Higher odds of outcome
(Age Group) 55-64	1.504686845	0.504686845	Higher odds of outcome
Preceding Presidential	1.49462119	0.49462119	Higher odds of outcome
(Age Group) 45-54	1.394386969	0.394386969	Higher odds of outcome
Indian-American	0.743525891	0.256474109	Lower odds of outcome
(Race) Two or More	0.743892057	0.256107943	Lower odds of outcome
Midterm_Pct_Over25	1.255983712	0.255983712	Higher odds of outcome
(Party) Republican	1.250974903	0.250974903	Higher odds of outcome
(Age Group) 35-44	1.229512698	0.229512698	Higher odds of outcome

TABLE 5

glm variable importance	
	Overall
Total Vote	284.42
Preceding Presidential	69.69
(Age Group) 55.64	53.44
(Age Group) 45.54	45.54
Midterm_Pct_Over25	40.62
(Party) Republican	36.63
(Age Group) 35.44	28.70

Appendix A: Correlation Table

Correlation between census variables determined by built-in R function cor().Variables in blue were excluded in the model due to the high correlation with other variables. Reference variables are not shown.

	Pct_Under40	Pct_Under60	Pct_Above60	Pct_Hisp	Pct_White	Pct_Black	Pct_Asian	Pct_Owner	Avg_House_Inc	Avg_Inc	Avg_House_Size	Avg_House_Value	Pct_HS_Educ	Pct_BS_Educ	Pct_White_Collar	House_Inc_Below55K
Pct_Under40	1															
Pct_Under60	-0.8441	1														
Pct_Above60	-0.4713	0.2473	1													
Pct_Hisp	0.2486	-0.4639	-0.1980	1												
Pct_White	-0.4631	0.6271	0.4828	-0.6756	1											
Pct_Black	0.4326	-0.5971	-0.4007	0.5083	-0.9428	1										
Pct_Asian	0.1198	0.0368	-0.3474	-0.1876	-0.0363	-0.1856	1									
Pct_Owner	-0.8068	0.7543	0.1541	-0.3230	0.5021	-0.4709	-0.0661	1								
Avg_House_Inc	-0.4112	0.6436	0.1700	-0.6948	0.7803	-0.7953	0.3824	0.4853	1							
Avg_Inc	-0.1806	0.4556	0.2692	-0.6920	0.7730	-0.7708	0.3324	0.2230	0.9328	1						
Avg_House_Size	-0.5881	0.4077	-0.3013	0.1959	-0.1651	0.1093	0.0378	0.6627	-0.0448	-0.3841	1					
Avg_House_Value	-0.1243	0.2702	0.3924	-0.6029	0.7018	-0.6569	0.1364	0.0758	0.7680	0.8654	-0.4484	1				
Pct_HS_Educ	0.1653	-0.4301	-0.1208	0.6393	-0.7075	0.7514	-0.4574	-0.2819	-0.9148	-0.9272	0.2373	-0.7572	1			
Pct_BS_Educ	-0.0757	0.3406	0.1596	-0.6518	0.6784	-0.7142	0.4712	0.1188	0.8911	0.9506	-0.3812	0.8289	-0.9665	1		
Pct_White_Collar	-0.2169	0.4947	0.1265	-0.7340	0.7258	-0.7294	0.4235	0.3379	0.9353	0.9381	-0.2257	0.7681	-0.9669	0.9512	1	
House_Inc_Below55K	0.1290	-0.2535	-0.0728	0.3101	-0.4305	0.4363	-0.0995	-0.4359	-0.4292	-0.4125	0.0017	-0.3077	0.5121	-0.4005	-0.4980	1

Appendix B: The Interactive Model

The probability model uses the coefficients from the logistic output to calculate probabilities based on the standard formula:

$$\hat{p} = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p)}{1 + \exp(b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p)}$$

Where $\hat{p}$ is the expected probability, b_0 is the intercept, b_p are the variable coefficients, and X_p are the variable inputs.

Using the model:

The model can be found at: https://github.com/rtwrtw8/HVT/tree/master/Interactive%20Model. Columns G and H represent the model input of before and after for a particular voter. Users may modify any X_p in the model to see the change in probability of that voter turning out. For example, setting cell G5 = 0 and H5 = 5, one may observe the change in probability of a voter turning out when his or her total vote changes from 0 to 5 (assuming every other variable is identical across the columns G and H). The probability is calculated and shown at the bottom of the sheet, in cells G37 and H37 respectively, with the change in probability shown in H38. The 'Reset' button at the top will reset the inputs for all the variables under column G that were arbitrarily picked to be reset to, *except for the census variables (rows 24-34 inclusive)*: these variables are reset to the state averages.

There are restrictions to the inputs one may provide for X_p, which is built in the model and listed as follows:

1. The total vote is bucketed at five levels; the valid input value range is [0, 5]
2. For all the binary variables, the valid input value is 0 or 1
3. For the percentage variables, the valid input value is between 0 and 100% (The model estimates in the interactive model have been adjusted for the input variable scale – from the percentage units to fractions of 1).
4. User must not enter multiple 1s for each category "block" as a voter cannot be registered, for example, as a male AND a female in the voter files. Same rule applies to all the census variables – the percentages for each category block cannot exceed 100%. If there is 60% of the population between the ages of 20 and 40 in a neighborhood, there cannot be 70% of the population between the ages of 40 and 60 in the same neighborhood.

Rules # 1 – 3 are controlled by data validation and rule #4 is controlled by the "Calculation" button in the model.

Users must also keep in mind that the reference variables are still integrated in the model even though they are not shown. For example, if Libertarian, Republican, and Affiliated are all 0, then Democrat (which was the reference variable for party, as mentioned above) is automatically treated as a 1.