­­­­­­­­­­Independent Project

Data Analysis for the Social Sciences

Prof. Gregory Eirich

Yun Choi

**Is Vote-By-Mail A Partisan Matter?: Lessons from linking NC elections data**

1. **Research Topic**

Throughout the 2020 U.S. presidential election, then-President Donald Trump continued his attacks on mail-in voting. Although the extent of his influence remains unclear, only 30% of Republicans voted by mail despite nationwide polling place closures and the health risk of casting an in-person ballot during the COVID-19 pandemic. In comparison, nearly 60% of Democrats cast a mail-in ballot. Such a disparity raises the question of whether vote-by-mail (VBM) is indeed a partisan issue and less accepted by Republicans, even when controlling for other factors that influence one’s preference for VBM. Survey studies have found that Republicans are less likely to VBM than Democrats, but they did not consider other factors that might influence their preference of voting method.

1. **Hypothesis**

In this paper, I examine the following hypothesis: (1) being a registered Republican (independent variable) decreases the probability of VBM (dependent variable), even when controlling for voter age, race, sex, and distance to a designated polling location (controlled variables); and (2) being white has differential impacts on the probability of VBM across the political spectrum. I hypothesize that within the Republican Party, being white intensifies disapproval against VBM. On the other hand, within the Democratic Party, being white increases the probability of VBM.

I control for voter age and distance to a polling place because age and travel distance are two primary factors in deciding whether to vote-by-mail (VBM). The older the voter, and the longer the travel distance, the higher the voter's incentive to VBM and save the trip. I also control for voter race and gender because white, and male voters are likely to have a higher socioeconomic status than non-white, and female voters, respectively, which is also one of the strong predictors for a high probability of VBM.

My first hypothesis will be falsified if the probability of VBM is constant across the political spectrum, net of all the controlled variables. That means the disparity in the probability of VBM derives from some or all of the variables I control for. My second hypothesis will be falsified if being white influences the probability of VBM in the same direction across the political spectrum. That means being white has the same effect on the probability of VBM, regardless of their political affiliation.

1. **Data Description**

My ideal data would show each voter's demographic and socioeconomic characteristics, political affiliation, and preference for the voting method. However, few states have made such personal information public due to privacy reasons. I have chosen North Carolina among all states because its voter file has individual-level demographic information such as race, ethnicity, and sex, unavailable in most states. And it is open to the public.

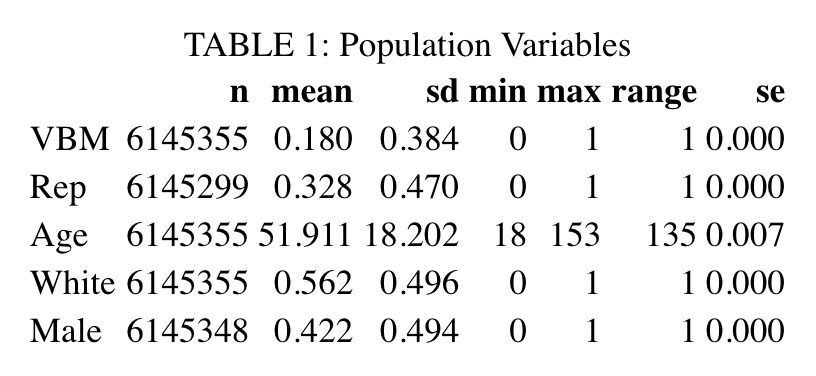
By linking various elections datasets – voter registrations, voter history, and polling places – from the North Carolina State Board of Elections, I have created a dataset that shows the demographic, political, and voting preference information for each voter in North Carolina. I filtered 6.1 million who voted for the 2020 presidential election only, and randomly selected 10 samples from each of nearly 2,600 precincts in North Carolina. That way, all precincts in different parts of the state are represented. That has left me with about 26,000 samples in total, which is about 0.4% of the entire 2020 presidential voters in North Carolina.

I focus on the 2020 presidential election because during the period, VBM was most needed due to the global pandemic, but also caused a strong conflict between Democrats and Republicans. I selected a subset of voters due to the limits of computing power required for geocoding. It was infeasible to geocode residential and polling location addresses and calculate the distance between the two points for each voter.

Below is the list of variables I use in this paper:

* **Dependent:** VBM (binary - '1' if voted by mail; '0' if not) I created a new dummy variable showing whether each voter voted by mail during the 2020 presidential election. Those who voted by mail have a ‘1’ as the value, while the others have a ‘0’.
* **Independent:** Rep (binary - '1' if registered Republican; '0' if not) I created a new dummy variable showing whether each voter was a registered Republican, based on voter party affiliation status during the election. Registered Republicans have a ‘1’ as the value, while the others have a ‘0’.
* **Controlled 1:** Age (continuous - age as of 2020) I created a new continuous variable based on the date of birth column in the voter registration data. I calculated the voter age as of the date of the 2020 presidential election.
* **Controlled 2:** White (binary - '1' if non-Hispanic white; '0' if not) I created a new dummy variable that shows whether the voter is non-Hispanic, based on two preexisting columns – race and ethnicity. Non-Hispanic whites have a ‘1’ as the value, while the others have a ‘0’.
* **Controlled 3:** Male (binary - '1' if male; '0' if not) I created a new dummy variable that shows whether the voter is male, based on the preexisting ‘sex’ column. Males have a ‘1’ as the value, while the others have a ‘0’.
* **Controlled 1:** TravelDistance (continuous – travel distance b/w residence and polling location in miles) I geocoded each voter’s residence and designated polling place and calculated the geodesic distance between two points using the ‘gdist’ function in the ‘Imap’ package in R because of the ease of its calculation. The function uses the Vincenty inverse formula for ellipsoids. However, the geodesic distance would differ from actual travel distance over a road network. I intend to recalculate the travel distance between the two points using the ‘gmapdistance’ function in the identically named R package. The function uses the Google Maps Distance Matrix API to compute the distance between two points in meters.
* **Interaction:** Rep \* White (binary – ‘1’ only if non-Hispanic white Republican)

1. **Descriptive Statistics**

****

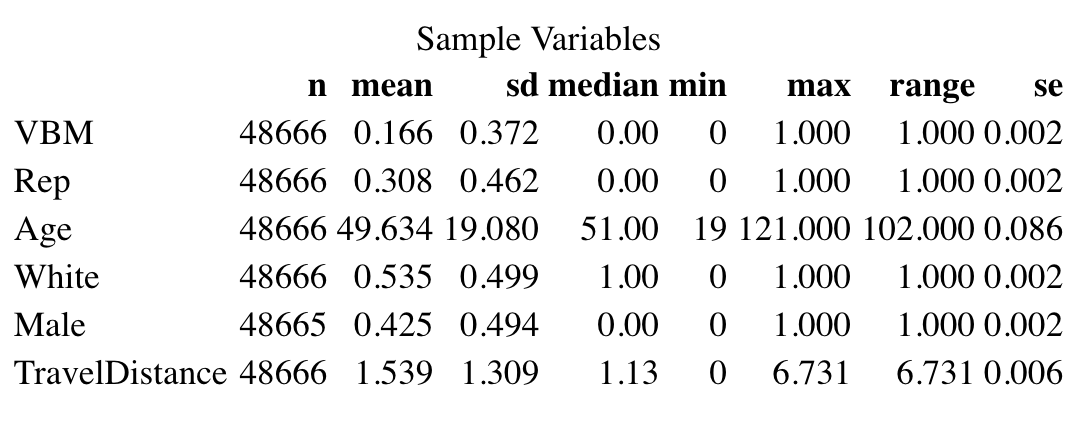
****

TABLE 1 shows the mean, standard deviation, and range for each variable for the population (6.1 million North Carolinians voted for the 2020 presidential election) and the sample. The comparison between the two tables shows that the sample is representative of the population. TABLE 1 shows that the population is 33% Republican, 56% non-Hispanic white, 43% male, and have an average age of 52. And 18% of the population voted by mail. The sample is 31% Republican, 54% non-Hispanic white, 42% male, and have an average age of 50. And 17% of the sample voted by mail.

Although the sample is representative of the population, I find disparities between the statistics from the state voter file and other external data. The 2019 Census 1-year population estimates show that North Carolina’s total population is 62.5% non-Hispanic white, while TABLE 1 shows that North Carolina’s voting population is 56% non-Hispanic white. Moreover, presidential election voting history shows that North Carolina has voted Republican in nine of the last 10 presidential elections. That implies many North Carolinians who actually vote for the Republican Party are not registered as Republicans.

For this paper, I disregard these questions due to time constraints, but I intend to explore these issues later.

1. **Initial Models**

I run a logistic regression with the observational final dataset created by linking various elections datasets from North Carolina. I choose logistic regression because it allows me to work with a binary dependent variable and easily separate the effects of other factors that influence the dependent variable. Whether to VBM is a decision influenced by interactions between various voter characteristics. Therefore, it is necessary to separate the effects of them to measure the sole effect of being Republican on one's probability of VBM.

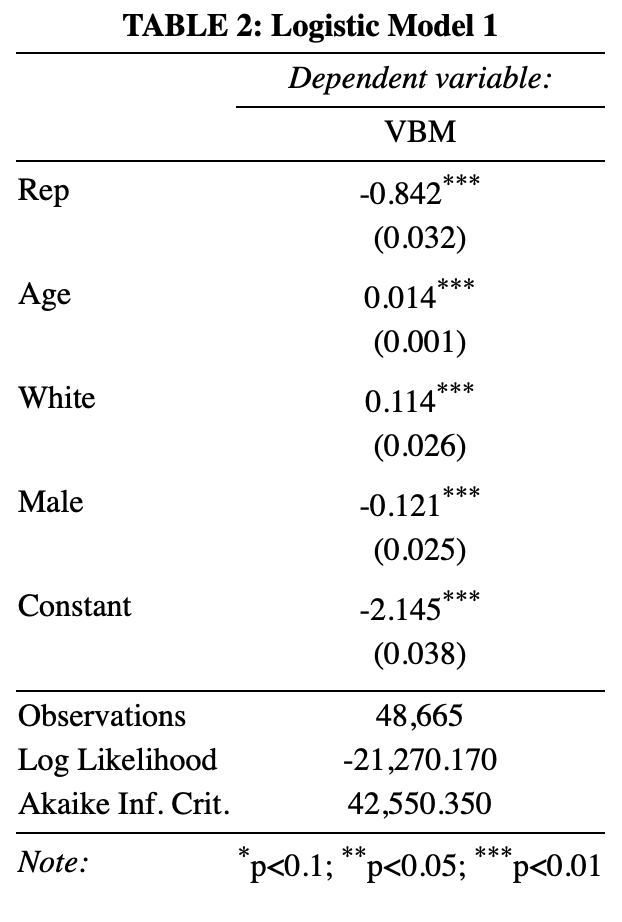
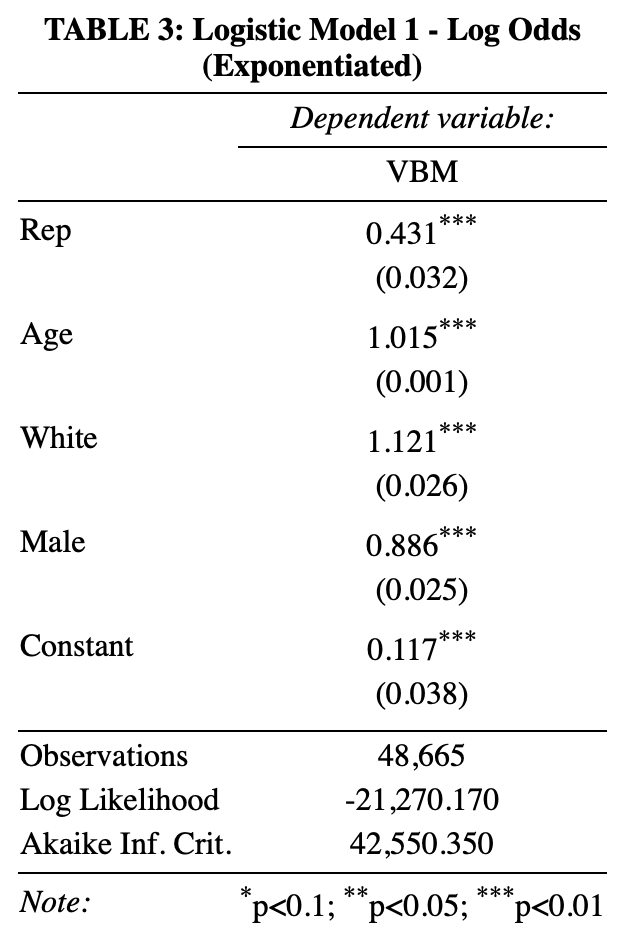
I have chosen observational data over survey data because the latter does not have information on each respondent’s residential address and polling location, precluding the calculation and incorporation of the travel distance to a polling location, which is a decisive factor in determining whether to VBM. Voters assigned to a polling location far from their residence have a higher incentive to vote by mail than in person. Therefore, incorporating the travel distance variable in the model is essential.

Logistic regression assumes (1) linearity in the logit for continuous variables, (2) sample representativeness, (3) absence of multicollinearity, (4) lack of strongly influential outliers, and (5) independence among observations. The first four conditions can be easily met. First, there is no reason to assume non-linearity in the logit for the continuous variables in the model – age, travel distance, and total household income estimate. Second, the individual-level voter file contains information for the entire voter population in North Carolina. Third, no two or more independent variables are strongly correlated to the level where they cannot provide unique information. Fourth, even if several outliers exist, their influence will be minimal due to the large sample size.

However, the last assumption – independence among observations – can be violated because a person’s decision to vote in person on Election Day is likely to influence the decisions of their family members. For example, a father’s decision to vote in person on Election Day is likely to influence others to join in-person voting by drastically reducing the cost of traveling to the location for the whole family. His family members can simply hop on his car to get to the polling place. The travel can be joyful and even lead to an exciting family outing.

However, this issue should have been resolved in the sampling process where I randomly selected 10 samples from each precinct. It is very unlikely that two members from one family got selected in the process.

* 1. **Multiple Logistic Model without TravelDistance Variable**

****

My initial logistic regression model has four explanatory variables: Rep, Age, White, and Male. In TABLE 2, the coefficient of -.842 on Rep means that for being registered Republican, the log odds of VBM decreases by .842 on average, controlling for voter age, race, and sex. The coefficient of .01 on Age means that for each year increase in age, the log odds of VBM increases by .01 on average, controlling for voter race, sex, and political affiliation. The coefficient of .11 on White means that for being white, the log odds of VBM decreases by .11 on average, controlling for voter age, race, and sex. The coefficient of -.12 on Male means that being male, the log odds of VBM decreases by .12.

TABLE 3 shows the odds for each variable. Coefficients in logit models are odds taken a natural logarithmic function. To retrieve actual odds, coefficients in logit models should be exponentiated and subtracted by one.

However, odds, the probability of an event occurring divided by the probability of the event not occurring, is not very intuitive. Calculating the Odds-Ratio (OR), the odds of the event in one group exposed to a treatment, divided by the odds in another group not exposed, can give more interpretive results. An OR can calculated by calculating the difference between the odds and the number one.

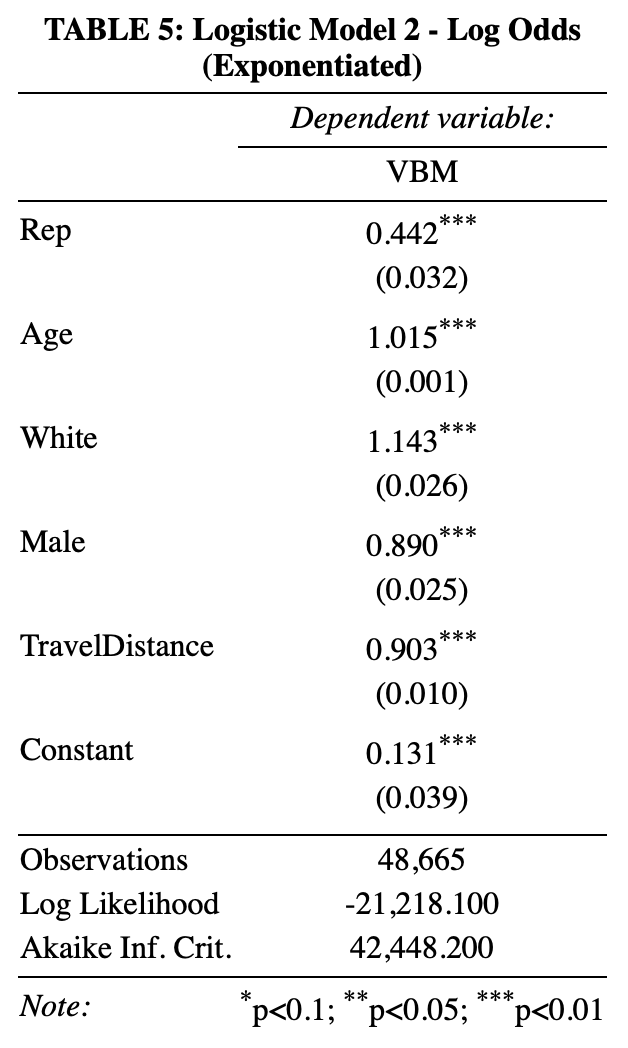
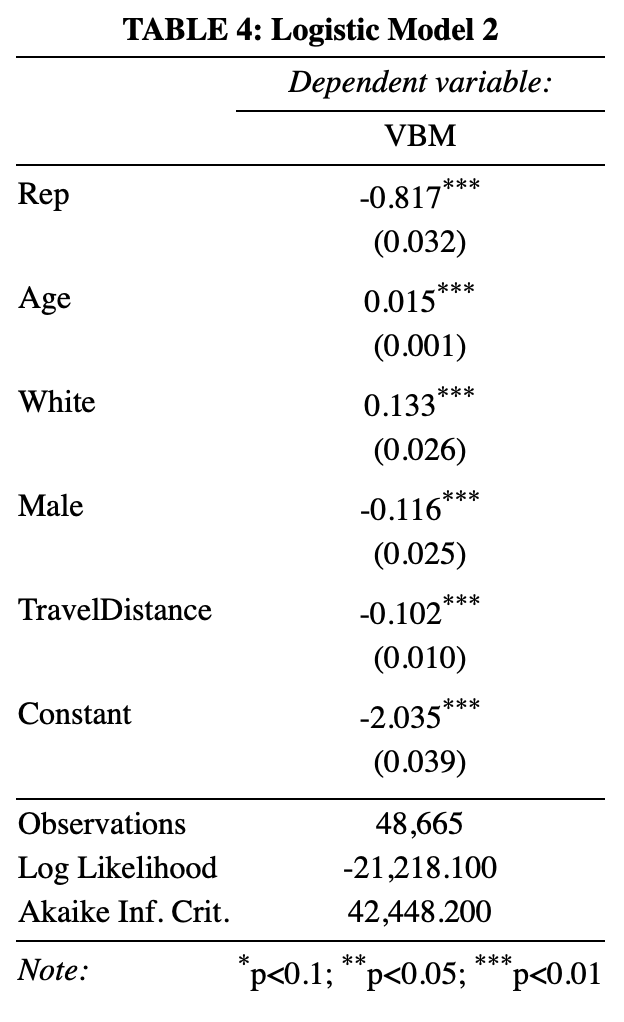
An Odds-Ratio (OR) of 1 means that given a one-unit change in that independent variable, there is an equal chance of the event happening as not happening. An OR above 1 indicates that given a one-unit change in that independent variable, there is a lower chance of the event happening than not happening. An OR below 1, on the other hand, indicates that given a one-unit change in the independent variable, there is a higher chance of the event happening than not happening.

TABLE 3 shows the odds of .43 on Rep. That means for Republicans, the odds of VBM decrease by 57% (.43-1 = -.57) on average, compared to non-Republicans with the same age, race, and sex. The odds of 1.01 on Age means that the odds of VBM increase by 1% (1.01-1=.01) on average, compared to a voter who is a year younger but with the same race, sex, and political affiliation. The odds of 1.12 on White means that the odds of VBM increase by 12% (1.12-1=.12) on average, compared to a non-white voter with the same age, sex, and political affiliation. The odds of 0.89 on Male means that the odds of VBM decrease by 11% (.89-1 = -.11) on average, compared to a female with the same age, race, and political affiliation.

Though questionable, the p-values of the coefficients of all explanatory variables are statistically significant. The model has the log-likelihood value of -21,270, and the Akaike Information Criterion (AIC) value of 42,550. The log-likelihood and AIC values from the initial model do not explain anything about the goodness of fit because they become meaningful only when compared to the equivalent values from a similar model with different specification. A higher log likelihood value indicates a better fit to the data. On the other hand, a higher AIC value indicates a worse fit to the data.

The results from my initial model confirm my first hypothesis that being Republican decreases the probability of VBM, controlling for other factors. However, this model does not factor in the cost of traveling, which is one of the primary factors in deciding whether to VBM. Therefore, I add TravelDistance variable in the next model.

* 1. **Multiple Logistic Model with TravelDistance Variable**



My second logistic regression model has one more explanatory variable compared to the initial model: TravelDistance. Adding the new variable has slightly changed the coefficients on the pre-existing terms, which means that some of the variance explained by the pre-existing variables in the first model can actually be explained by the new TravelDistance variable.

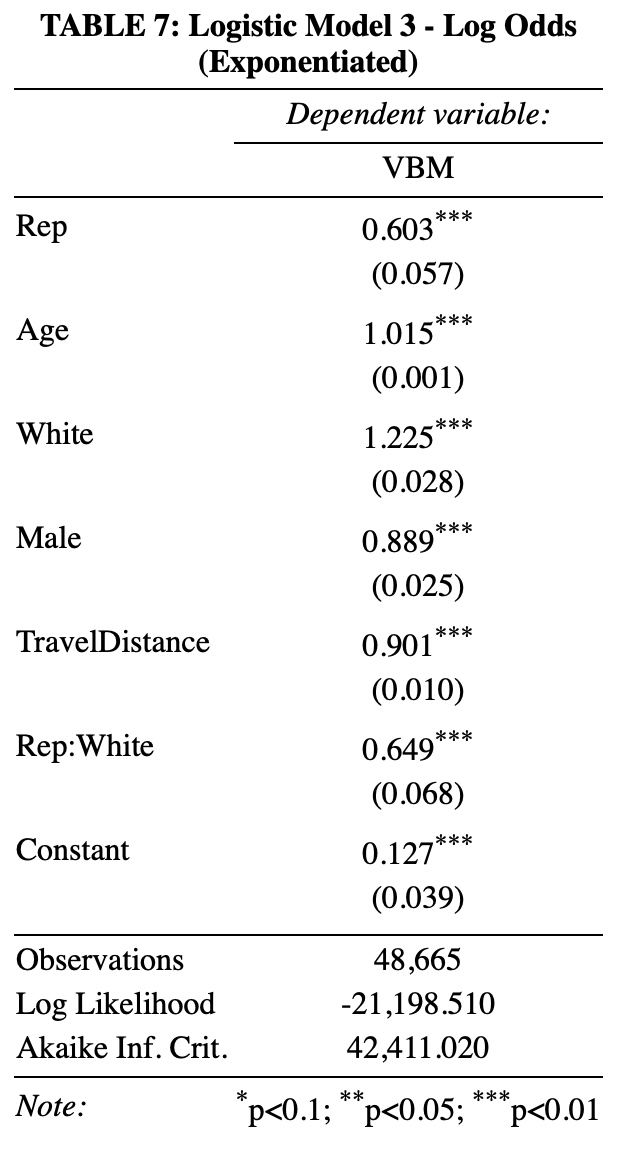
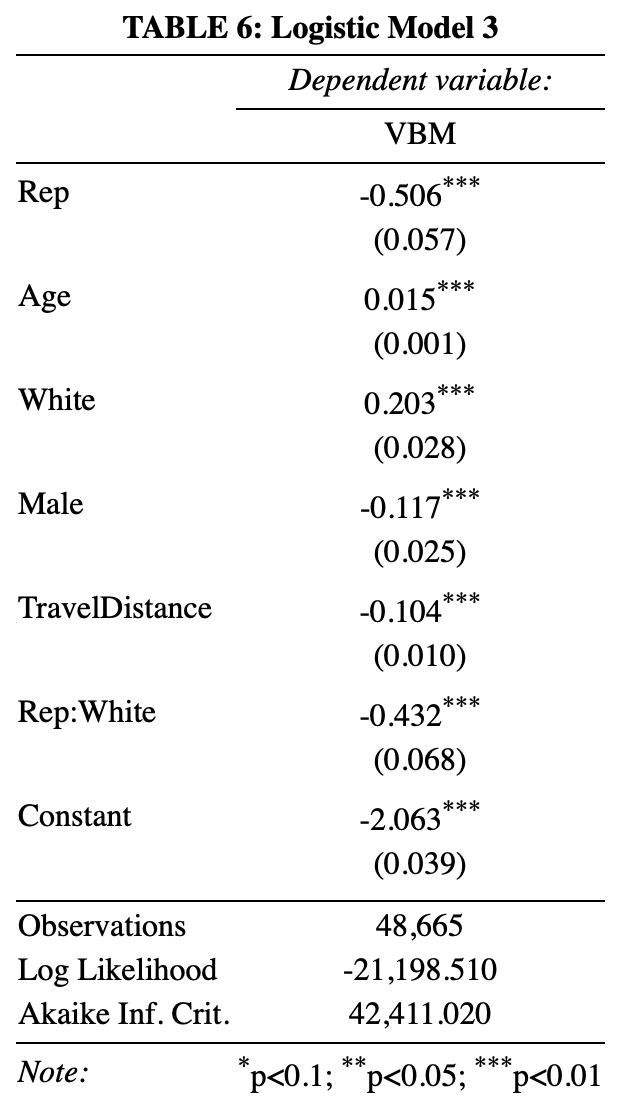
Compared to the coefficients in the initial model, TABLE 3 shows that the absolute value of the coefficient on Rep has decreased from .84 to .82. That means a small part of the disparity in the probability of VBM between Republicans and Democrats can be explained by the difference in the average distance to a designated polling place. On the other hand, the coefficients on Age and White have increased from .01 to .02, and from .11 to .13, respectively. The coefficient on Male has remained same. The coefficient on TravelDistance is -.1 and is statistically significant. That aligns with my expectation that the longer the distance to a designated polling place, the less likely the voter is to VBM.

TABLE 6 shows the odds of .44 on Rep. That means for Republicans, the odds of VBM decrease by 56% (.44-1 = -.56) on average, compared to non-Republicans with the same age, race, sex, and distance to a polling place. The odds of 1.02 on Age means that the odds of VBM increase by 2% (1.02-1=.02) on average, compared to a voter who is a year younger but with the same race, sex, and political affiliation. The odds of 1.14 on White means that the odds of VBM increase by 14% (1.14-1=.14) on average, compared to a non-white voter with the same age, sex, and political affiliation. The odds of 0.89 on Male means that the odds of VBM decrease by 11% (.89-1 = -.11) on average, compared to a female with the same age, race, and political affiliation.

Although the coefficients on explanatory variables have changed, their statistical significances remain below 0.01. The model’s log likelihood and AIC values indicate that adding DistanceTravel variable was the right choice. The model has a log likelihood value of -21,218 and an AIC value of 42,448. That is a slight improvement from the previous model with a log likelihood value of -21,270, and an Akaike Information Criterion (AIC) value of 42,550. A higher log likelihood value and a lower AIC value indicate a better fit to the data.

However, this model is unable to capture the interaction between Rep and White variables, which is crucial in proving my second hypothesis on the differential impact of being white on the probability of VBM across the political spectrum. Therefore, in the next model, I add the interaction term between Rep and White to the current model.

* 1. **Multiple Logistic Model with TravelDistance Variable**



The addition of the interaction term between Rep and White has drastically changed the coefficients on Rep and White. The coefficient on Rep has dropped from .82 to .51, while the coefficient on White has increased from .13 to .2. However, the coefficients cannot be interpreted directly because of the interaction term.

Due to the interaction term, the coefficient on Rep changes depending on voter race. If a voter is not white (White = 0), the coefficient is -.51. If a non-Hispanic white voter (White = 1), the coefficient is subtracted by -.43. to -.94.

Among non-white voters (White = 0):

Among white voters (White = 1):

The results above mean that among voters of color, being Republican lowers the log odds of VBM by .51 on average, controlling for age, sex, and distance to a designated polling place. Among non-Hispanic whites, being Republican lowers the log odds of VBM by -.94 on average, controlling for age, sex, and distance to a polling place.

The coefficient on White also changes depending on voter political affiliation. If a voter is not Republican (Rep = 0), the coefficient is .2. If Republican (Rep = 1), the coefficient is added by -432, to -.23.

Among non-Republican voters (Rep = 0):

Among Republican voters (Rep = 1):

Among non-Republicans (Rep = 0), the coefficient of .2 on White means that for being white, the log odds of VBM increases by .2 on average, controlling for voter age, race, and sex. Among Republicans (Rep = 1), the coefficient of -.23 on White means that for being white, the log odds of VBM decreases by .23 on average, controlling for voter age, race, and sex.

Exponentiating the coefficients on Rep and White gives the following results:

Above odds can be interpreted as below:

* Among **non-white voters**, for being Republican, the odds of VBM decrease by 40% (.6-1 = -.40) on average, compared to a non-Republican with the same age, race, sex, and distance to a polling place.
* Among **white voters**, for being Republican, the odds of VBM decrease by 61% (.39-1 = -.61) on average, compared to a non-Republican with the same age, race, sex, and distance to a polling place.
* Among **non-Republican voters**, for being white, the odds of VBM increase by 22% (1.22-1=0.22) on average, compared to a non-white voter with the same age, sex, and distance to a polling place.
* Among **Republican voters**, for being white, the odds of VBM increase by 21% (.79-1=0.21) on average, compared to a non-white with the same age, sex, and distance to a polling place.

The coefficients and odds on the remaining variables do not change regardless of voter race or political affiliation. In TABLE 6, the coefficient of .02 on Age means that for each year increase in age, the log odds of VBM increases by .02 on average, controlling for voter race, sex, political affiliation, and distance to a polling place. The coefficient of -.12 on Male means that being male decreases the log odds of VBM by .12 on average, controlling for voter age, race, political affiliation, and distance to a polling location. The coefficient of -.1on TravelDistance means that for each mile increase in the travel distance to a polling place, the log odds of VBM decreases by -.1 on average, controlling for age, race, sex, and political affiliation.

In TABLE 7, the odds of 1.02 on Age means that the odds of VBM increase by 2% (1.02-1=.02) on average, compared to a voter who is a year younger but with the same race, sex, and political affiliation. The odds of .89 on Male means that the odds of VBM decrease by 11% (.89-1 = -.11) on average, compared to a female with the same age, race, and political affiliation. Lastly, the odds of .9 on TravelDistance means that the odds of VBM decrease by 10% (.9-1=-.1) on average, compared to a voter who has their polling place one mile further away, but with the same age, race, sex, and political affiliation.

However, it is important to separate probability and odds because the odds can have a large magnitude even if the underlying probabilities are low, and vice versa. Setting the voter age, sex, and distance to a polling place as 38 (the median age of the U.S. population), male, and one mile, respectively, produce more intuitive results.

* The comparison of 38-year-old non-white vs. white Republican male voters shows that:
  + The predicted probability of VBM for a 38-year-old Republican male of color is .10, which is 10%.
  + The predicted probability of VBM for a white Republican male ages 38 is .08, which is 8%.
* The comparison of 38-year-old non-white vs white non-Republican male voters shows that:
  + The predicted probability of VBM for a 38-year-old non-Republican male of color is .15, which is 15%.
  + The predicted probability of VBM for a 38-year-old white non-Republican male is .18, which is 18%.

TABLE 8: Predicted Probability of VBM among 38-year-old male voters who live a mile away from their polling place.

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted Probability  to VBM | Non-white | White | Difference  (White - Non-white) |
| Republican | 10% | 8% | **-2%** |
| Non-Republican | 15% | 18% | **+3%** |
| Difference | 5% | 10% |  |

TABLE 8 shows that being white has differential impact on the probability of VBM across the political spectrum. Among Republicans, being white decreases the probability of VBM by 2% points. On the other hand, among Democrats, being white increases the probability of VBM by 3% points.

Although the coefficients on explanatory variables have drastically changed, as TABLE 6 shows, their statistical significances remain below 0.01. The coefficient on the new interaction term is substantially and statistically significant, which signifies adding the interaction term was the right choice. The model has a log likelihood value of -21,199 and an AIC value of 42,411. That is a slight improvement from the second model with a log likelihood value of -21,218, and an Akaike Information Criterion (AIC) value of 42,448. A higher log likelihood value and a lower AIC value indicate a better fit to the data.

Above results confirm both of my hypotheses. The model’s negative coefficient on Rep means that being Republican lowers the probability of VBM, controlling for other voter characteristics, such as age, race, sex, and distance to a polling place. Being white has a differential impact on the probability of VBM across the political spectrum. Within the Republican Party, being white decreases the probability of VBM. Conversely, within the Democratic Party, being white increases the probability of VBM.

1. **Final Model**

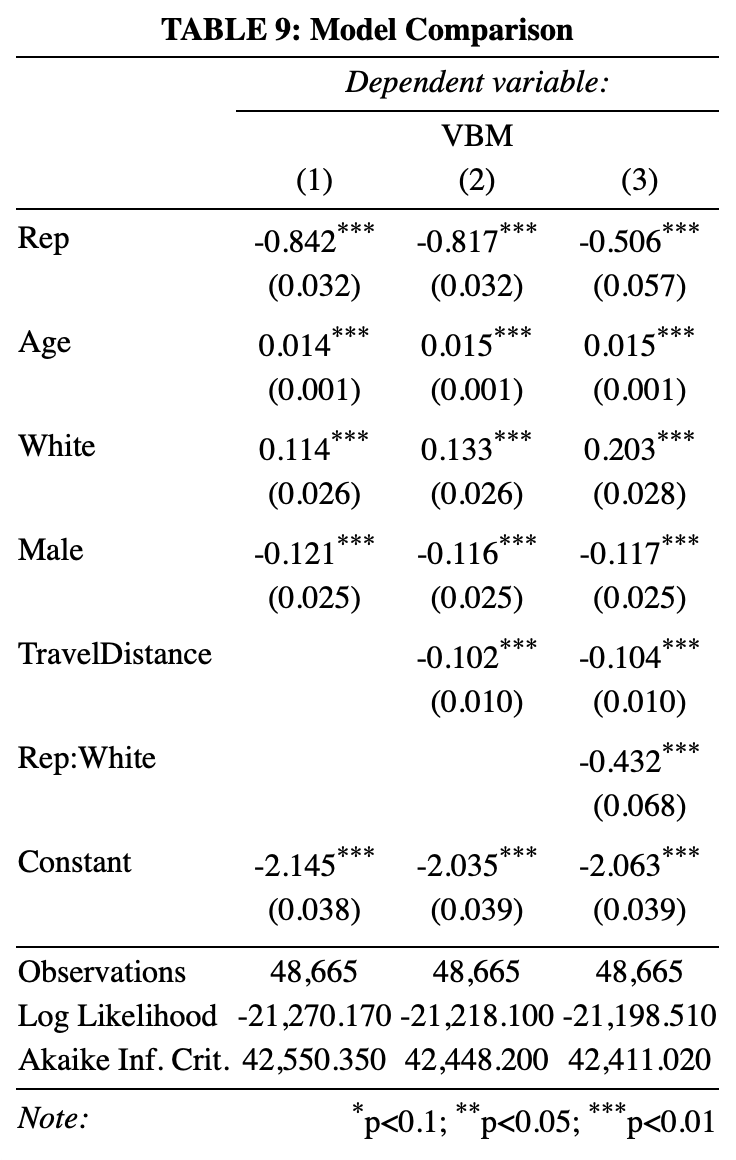


TABLE 9 shows the coefficients of the three models I compared in the above section. The first model had four explanatory variables – Rep, Age, White, and Male; the second model had one more variable, TravelDistance; and, the last model had an additional interaction term between Rep and White. Adding the TravelDistance variable to the first model did not change the coefficients on the variables as much as I had expected. However, adding the interaction term to the second model has greatly changed the coefficients on Rep and White.

I choose the third one as my final model based on theory and diagnostics. Theoretically, it is more natural to consider distance to polling place when choosing a voting method. Even if VBM is a partisan issue and strongly rejected by the Republican Party, a Republican who lives 10 miles away from their polling place would not want to take the trip, to exaggerate. Also, in real life, I see many leftist and rightest leaders are white. That implies being white might have a differential impact on the probability of VBM. Being white Democrat might increase the probability of VBM, while being white Republican might decrease the probability to VBM.

Statistically, TABLE 9 shows that the log likelihood and AIC scores have continuously improved as the model develops. The likelihood Ratio (LR) test results in TABLE 10 and TABLE 11 also show a statistically significant improvement between the first and second, and the second and third models, respectively.

TABLE 10: Likelihood Ratio Test between Model 1 and Model 2

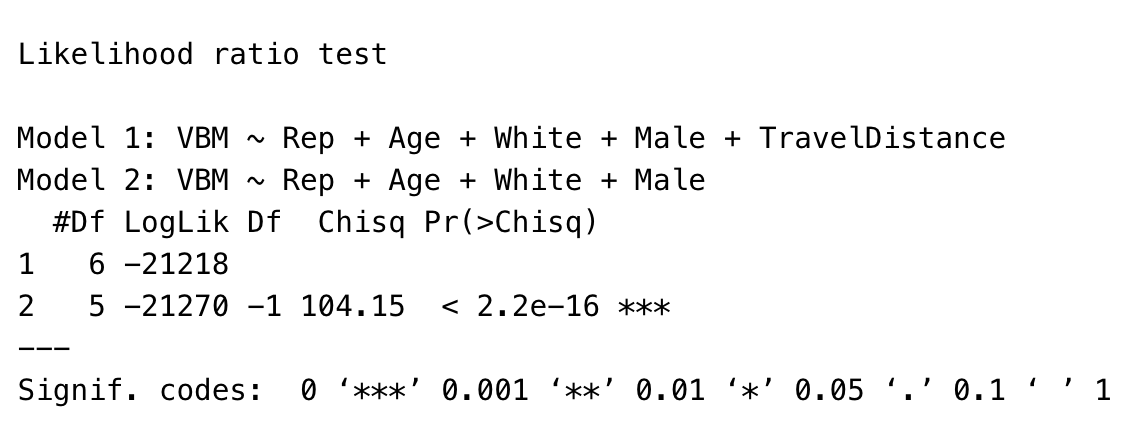
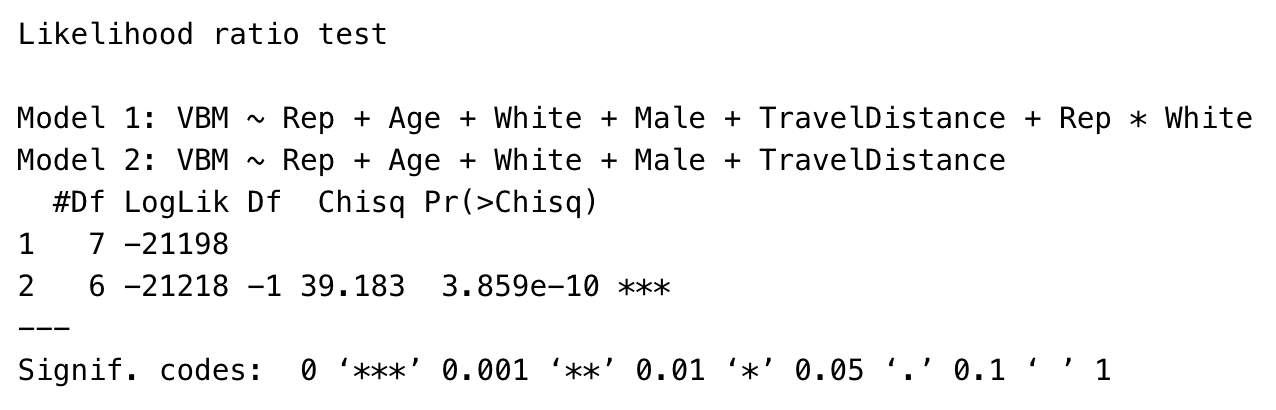


TABLE 11: Likelihood Ratio Test between Model 2 and Model 3



1. **Conclusion**

The results above support both of my hypotheses that (1) being Republican decreases the probability of VBM, controlling for the demographic and economic factors, such as voter age, race, sex, and distance to a polling place; and that (2) being white has a differential impact on the probability of VBM across the political spectrum. Being white increases the probability to VBM within the Democratic Party, and decreases the probability to VBM within the Republican Party.

Although my final model supports my hypotheses, there is still large room for improvement. I intend to improve this paper for the remaining semesters and submit it as a writing sample for PhD applications. Below is a list of improvements I would like to make:

First, I will change the way I calculated the travel distance. I used the geodesic distance because of the ease in calculation. However, the geodesic distance differs from actual travel distance over a road network. For a more precise calculation of the model coefficients, I will change the way I calculate the travel distance.

Second, I will add a new Income variable. I have intended to assign a median household income estimate to each voter by joining my dataset with census income data. However, I was unable to join a voter-level dot spatial dataset with the census tract-level income data, due to technical challenges and time constraints. Although this approach assumes the homogeneity in household income within each census tract, adding Income variable to the model would help produce coefficients closer to the truth, because income could be one of the strongest predictors for the probability of VBM.

Third, I will try imputing missing race and ethnicity data. From two missing data scholars who visited this semester’s Research Seminar class – Naijia Lu, and Rob Trangucci – I have learned that traditional missing data handling approaches that assume missing at random, such as Listwise Deletion, or Multiple Imputation, can bias results. Although I am currently not sure how to impute the missing data in voter file, I will review missing data literature, find the most appropriate approach to impute the missing data in voter file, try that approach, and see how my results change. I might find an interesting voting pattern among those who are less likely to fill out the voter registration form.