Analyzing Voter Behavior: A Logistic Regression Approach to Poll Data Interpretation*

Yingqi Pang

18 March 2024

Contents

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## # A tibble: 8 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   -0.291      0.222    -1.31    0.190
## 2 white          0.269      0.180     1.50    0.134
## 3 hispanic      -0.0359     0.236    -0.152   0.879
## 4 black         NA         NA        NA      NA
## 5 democrat       0.111      0.185     0.602   0.547
## 6 republican     0.126      0.203     0.618   0.536
## 7 independent   NA         NA        NA      NA
## 8 vote_history  -0.0298     0.137    -0.218   0.827
```

Introduction

In the realm of political polling, divergent methodologies can yield varying interpretations from the same dataset. This report presents a logistic regression analysis to predict the likelihood of voting for a particular candidate based on demographic factors and voting history. The model's framework is inspired by an experiment conducted by Nate Cohn in 2016, where raw poll data led to different results among pollsters due to methodological differences.

*Code and data are available at: https://github.com/pangyin2/cohn-model.git

Data

The dataset for this study was simulated to represent a sample of 867 likely voters with attributes including race, party affiliation, and voting history. Logistic regression was chosen for its appropriateness in binary outcome scenarios. Dummy variables were created for categorical attributes, and a model was fitted to ascertain the predictors' influence on voting for the Democratic candidate.

Hypothetical Analysis Based on the Model Output:

After fitting the logistic regression model, the summary(model) function will give us an overview of the model's coefficients, standard errors, z-values, and p-values.

Intercept: The intercept represents the log odds of voting for Clinton when all predictor variables are zero (which is not possible in this context since they are dummy variables). But technically, it's the baseline against which the effect of all other variables is measured.

Race: The coefficients for white, hispanic, and black would indicate the change in the log odds of voting for Clinton for each racial group compared to the base case (possibly another racial group not included in the model due to multicollinearity reasons).

Party Affiliation: Similar to race, the coefficients for democrat, republican, and independent will reveal the log odds of a Democrat, Republican, or Independent voting for Clinton compared to the baseline group.

Vote History: The coefficient for vote_history indicates the impact of having a voting history on the likelihood of voting for Clinton. A positive coefficient means that those with a vote history are more likely to vote for Clinton, while a negative coefficient suggests the opposite.

Results Analysis

The logistic regression output indicates the influence of each predictor on the likelihood of a vote for the candidate. Coefficients for race and party affiliation reveal their respective impacts when compared to their reference groups. The vote history's coefficient reflects the historical participation effect on the current voting decision. Significance was adjudged based on p-values, with a conventional alpha level of 0.05.

Discussion

Interpretation of coefficients suggests how each demographic factor contributes to voting behavior. A positive coefficient signifies an increased probability of voting for the candidate, whereas a negative one suggests the opposite. It's imperative to note the limitations inherent in the analysis due to the hypothetical nature of the data, which may not capture complex interactions or real-world polling nuances.

Conclusion

The analysis underscores the variability that can arise in polling outcomes based on methodological approaches. The logistic regression model provides a simplified yet insightful glance into how demographic and historical factors can sway electoral predictions. Real-world application would demand a nuanced approach, considering actual data and potentially more sophisticated statistical techniques.