

Biden to win the popular vote in the 2020 American presidential elections

Anne Collins, Jennifer Do, Andrea Therese Javellana, Wijdan Tariq

November 02, 2020

Abstract

“Why do/should we care about the problem and the results. What problem are we trying to solve. How did we go about solving or making progress on the problem? What is our answer? What are the implications of our answer?”

Keywords: Forecasting; US 2020 Election; Trump; Biden; Multilevel regression with post-stratification

1 Introduction

Forecasting is an integral human activity (Silver, 2012; Tetlock and Gardner, 2016) and voting is one of democracy’s most important civic duties. Putting the two together, therefore, forecasting the outcome of the presidential elections in the United States of America—one of the most important democracies on the planet—is perhaps the most widely discussed prediction in all of social discourse (after the weather, of course!).

In 2008, 89% of Americans had said that they read about the latest polls in the presidential contest (Erikson and Tedin, 2015) and it is expected that that number is even higher in 2020. There are important practical considerations as to why forecasting the presidential elections is important. It is reasonable to assume that the identity of the president of the United States has an influence on the likelihood that certain state policies may be adopted during the tenure of their presidency. Thus, stakeholders throughout the world would want to know the odds of certain policies being enacted in the future in order to be best prepared to mitigate or take maximum advantage when the time comes.

In this paper, we aim to predict the outcome of the overall popular vote of the 2020 American presidential election using public opinion polling data. People have different views over whether political polling are reliable predictors of election outcomes, especially after the shock victory of President Trump in 2016 which most pollsters failed to predict. Gelman and Azari (2017) point to nineteen lessons learned from that election, most relevant to our paper of which is the lesson that one needs to be cautious of survey nonresponse bias. Kennedy et al. (2018) showed that a late swing in vote preference toward Trump, a failure to adjust for overrepresentation of college graduates who mostly favored Clinton, and a clear change in voter turnout from 2012 to 2016 were some of the main reasons why pre-election polls performed poorly in 2016. Moreover, the rise of populism in America and the resulting countermovement makes elections more unpredictable than ever (Inglehart and Norris, 2016), even with the explosion of data availability. Notwithstanding these issues, the use of public opinion polling data continues to be common practice and we make use of it in this paper. (BRIEFLY discuss datasets here).

We find that Biden/Trump is likely to win! (briefly discuss results here).

The remainder of the paper is structured as follows. Section 2 discusses the datasets that we use and describes the data cleaning process. Section 3 introduces our models and discusses the multilevel regression with post-stratification methodology that we used. Section 4 presents our results. Finally, Section 5 discusses our results, addresses limitations, and suggests future avenues for work in this area.

(Introduce R, R packages used, logistic regression and MRP methods briefly here as well as a quick note on results..briefly introduce which data we use and what key findings were...WHO IS GOING TO WIN??...give a sketch of the rest of the report)

2 Data

Include a discussion on how the data was obtained and point out any biases if there are any. e.g. ‘the frame for the survey is all Xbox users so women are less likely to be sampled...’. We need to include a discussion like this about UCLA and ACS datasets.

3 Model

4 Results

5 Discussion, Limitations, and Future Work

Insert comment here from Kennedy and Gelman (2019, p. 19) about cost/benefits of MRP.

“Given the bias in the survey, discussed in Data section, we used MRP to. . . . While this would address some aspects of the bias [discuss bits that it can address], it cannot address all of it. This includes [discuss bits that it cannot address].”

Our model assumed people would either vote for Trump or Biden. In reality, independent voters would also. . . so. . . Multinomial logistic regression model would be more appropriate.

Due to the unavailability of variables in the post-stratification data, we were not able to use important survey questions pertaining to attitudes and behaviors of voters towards policies and contemporary issues. Policies that are of concern to voters are known to influence voter choice (Petrocik, 1996). Some beliefs and attitudes might be correlated with certain demographic variables which we used, and may have helped in improving the predictive power of our model. For example, it is known that economic perceptions among voters may be important predictors for election outcomes (Duch and Stevenson, 2008) but we were not able to include such variables in our model. We now know that negative sentiment toward Muslim Americans was a strong and significant predictor of supporting Trump in the 2016 presidential election (Lajevardi and Abrajano, 2019). Some psychological patterns have also been observed among voters (Womick et al, 2019). Future work and post-stratification data will hopefully make available such important attitudinal issues that could help improve model specifications.

Another thing to note is that in this digital age, dramatic shifts in outcomes are perhaps possible in a very short period of time, such as the period of time between when we tested our model and the day of the election. In particular, we are worried about how search engine manipulation can affect the votes of undecided voters (Epstein and Robertson, 2015).

Researchers have shown that while intention is the single best predictor of behaviour—in our case, the intention to vote for Trump or Biden—it is also important to take into account other factors such as environmental constraints and the skills necessary to perform the behaviour (Fishbein and Ajzen, 2011). We have not been able to take into account the impact of COVID-19 on the ability of certain segments of the American population to participate in the vote. In particular, our predictions do not address the impact of postal ballots.

Future work can explore ways in which social media data—which have been shown to be useful predictors of election outcomes (e.g. Burnap et al., 2016; DiGrazia et al., 2013; Tumasjan et al., 2010)—can be combined with the MRP methodology to derive even more powerful predictive models. Perhaps new and creative sampling methods may need to be established to ensure statistically reliable sampling when working with social media data (Metaxas et al. 2011).

References {indent: false}

Papers about US election predictions and predictions in general

- Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, 41, 230-233.
- Caughey, D., & Warshaw, C. (2017). Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014. *American Political Science Review*, 112, 2 (November 2017): 249–266.
- DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11), e79449.
- Duch, R. M., & Stevenson, R. T. (2008). *The economic vote: How political and economic institutions condition election results*. Cambridge University Press.
- Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33), E4512-E4521.
- Erikson, R. S., & Tedin, K. L. (2015). *American public opinion: Its origins, content and impact*. Routledge.
- Gelman, A., & Azari, J. (2017). 19 things we learned from the 2016 election. *Statistics and Public Policy*, 4(1), 1-10.
- Fishbein, M., & Ajzen, I. (2011). *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis.
- Inglehart, R. F., & Norris, P. (2016). Trump, Brexit, and the rise of populism: Economic have-nots and cultural backlash. HKS Working Paper No. RWP16-026.
- Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L. Witt, G. E., & Wlezien, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1), 1-33.
- Lajevardi, N., & Abrajano, M. (2019). How negative sentiment toward Muslim Americans predicts support for Trump in the 2016 Presidential Election. *The Journal of Politics*, 81(1), 296-302.
- Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011). How (not) to predict elections. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 165-171). IEEE.
- Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science*, 825-850.
- Silver, N. (2012). *The signal and the noise: why so many predictions fail-but some don't*. Penguin.
- Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In: *Fourth international AAAI conference on weblogs and social media*.
- Womick, J., Rothmund, T., Azevedo, F., King, L. A., & Jost, J. T. (2019). Group-based dominance and authoritarian aggression predict support for Donald Trump in the 2016 US presidential election. *Social Psychological and Personality Science*, 10(5), 643-652.
- Methodology papers*
- Alexander, R. (2020). Getting started with MRP. Teaching note.
- Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73(4), 641-678.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15), 2865-2873.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3), 762-776.

- Ghitza, Y., & Gelman, A. (2020). Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research. *Political Analysis*, 28(4), 507-531.
- Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis*, 27(2), 163-192.
- Hanretty, C. (2019). An introduction to multilevel regression and post-stratification for estimating constituency opinion. *Political Studies Review*, 1478929919864773.
- Jackman, S., Ratcliff, S., & Mansillo, L. (2019). Small area estimates of public opinion: model-assisted post-stratification of data from voter advice applications. Working paper.
- Kennedy, L. & Gabry, J. (2020). MRP with rstanarm. Available at: <https://cran.r-project.org/web/packages/rstanarm/vignettes/mrp.html>. 20 July 2020.
- Kennedy, L., & Gelman, A. (2019). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *arXiv preprint arXiv:1906.11323*.
- Lauderdale, B. E., Bailey, D., Blumenau, J., & Rivers, D. (2020). Model-based pre-election polling for national and sub-national outcomes in the US and UK. *International Journal of Forecasting*, 36(2), 399-413.
- Tausanovitch, C., & Warshaw, C. (2013). Measuring constituent policy preferences in congress, state legislatures, and cities. *The Journal of Politics*, 75(2), 330-342.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980-991.
- Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709-747.
- Packages used*
- Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2020). rmarkdown: Dynamic Documents for R. R package version 2.4, <https://github.com/rstudio/rmarkdown>.
- Arnold, Jeffrey B. (2019). ggthemes: Extra Themes, Scales and Geoms for ‘ggplot2’. R package version 4.2.0. <https://CRAN.R-project.org/package=ggthemes>
- Auguie, Baptiste (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. <https://CRAN.R-project.org/package=gridExtra>
- Firke, Sam (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. <https://CRAN.R-project.org/package=janitor>
- Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. <https://CRAN.R-project.org/package=stargazer>
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Wickham, Hadley (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Wickham, Hadley, et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
- Wickham, Hadley and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. <https://CRAN.R-project.org/package=scales>
- Xie, Yihui (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.
- Xie, Yihui (2019) TinyTeX: A lightweight, cross-platform, and easy-to-maintain LaTeX distribution based on TeX Live. *TUGboat* 40 (1): 30–32. <http://tug.org/TUGboat/Contents/contents40-1.html>