Biden to win the popular vote in the 2020 American presidential elections in a closely fought contest

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

November 02, 2020

Abstract

"The 2020 American presidential elections are the most discussed elections in the world due to their potential and profound impact on global affairs. In this paper, we aim to predict whether Biden or Trump will win the popular vote and by how much of a margin. We ran a logistic regression model using 5,127 observations from election survey data provided by Democracy Fund + UCLA Netscape. To provide more robust predictions, we utilize the multilevel regression with poststratification method using over 2 million observations from the American Community Survey (ACS) data. While the survey data suggests that Biden will get 52% of the popular vote compared to Trump's 48%, our post-stratification results suggest that the election will be even closer, with 51% predicted for Biden and 49% for Trump. We conclude the paper with a discussion of the benefits of using post-stratification on survey data to correct for potential sample biases and to get more reliable regression estimates."

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

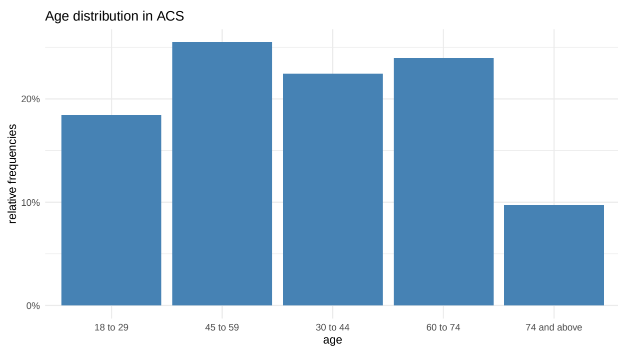
The survey dataset we used was retrieved from the Nationscape Data Set. This dataset was created in partnership between the Democracy Fund Voter Study Group and UCLA Political Scientists Chris Tausanovitch and Lynn Vavreck (Tausanovitch and Vavreck, 2020). Nationscape had conducted surveys to 500,000 Americans from July 2019 to December 2020 in lead up to the 2020 campaign and election. Each week, the survey team had interviewed roughly 6,520 people. To provide access to audiences, Lucid, a market research platform, had provided samples and an online exchange for survey respondents for Nationscape. The samples taken from this exchange included demographic quotas including age, gender, ethnicity, region, income and education. All respondents had conducted the survey online, along with an attention check prior to responding to the survey. The survey team interviewed people across the U.S, and had accounted for respondents in nearly all counties, congressional districts, and mid-sized U.S cities. To ensure accurate representation of the American population, the survey data was weighted and generated using a raking technique, with weights generated per week's surveys. The weights were derived from the 2017 American Community Survey of the U.S Census Bureau's adult population and its respective demographics (such as gender, region, race, household income, education, age, 2016 presidential vote, etc.). As well, representativeness was followed according to the Pew Research Center's evaluations of non-probability samples. The Nationscape results were compared to results from the 2018 and 2016 Pew Reports. It was determined that the Nationscape estimates were close to samples by the Pew Research Center.

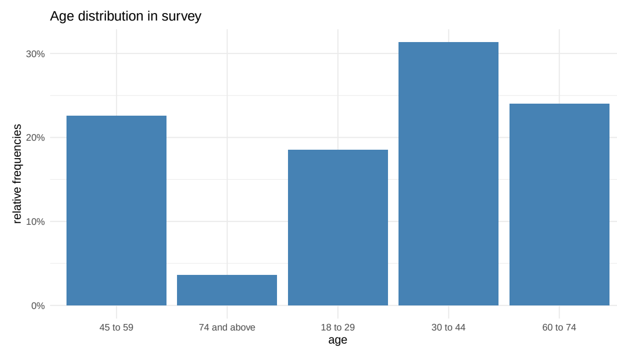
The data set we used was obtained from the Phase 2 of Nationscape's data, and accounted for the Nationscape Wave 50, taken from June 25th to July 1st, 2020. As the Nationscape data set had hundreds of variables, we chose to extract 11 variables focusing on demographics. These variables included: age, gender, employment, race, Hispanic ethnicity, household income, state, census region, education, nativity (where the respondent was born), and their choice in vote between Donald Trump and Joe Biden in the upcoming 2020 election. Variables such as race and Hispanic ethnicity were combined into a new race variable, to account for Hispanic ethnicity as a race. As well, responses to voter choice which did not include Trump or Biden as answers were deleted, to align with our primary goal of predicting the winner of the 2020 election between these two primary candidates. We had also created bins for variables race, education, and employment to match the cleaned ACS data set for post stratification (further discussed later on). From these variables, we chose to further analyze 5 of these variables (age, gender, employment, race, and voter choice) to coincide with the variables analyzed in the ACS data set.

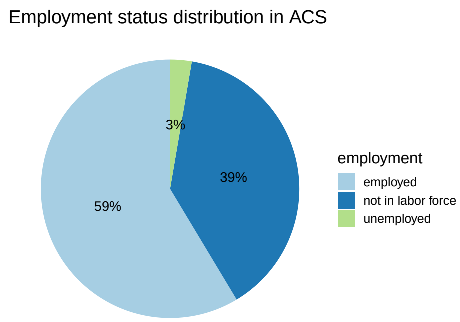
The data we used to post-stratify was the 2018 American Community Survey (ACS) data which we downloaded from the Integrated Public Use Microdata (IPUMS) US project website (Ruggles et al., 2020). The ACS is a national survey that is conducted annually. Participation in the survey is manadatory by law. It supplements the census and provides annual data on information to determine federal and state funds in America (the target population). The U.S. Bureau contacts 3.5 million randomly selected American households (the sampling frame) from their master address file each year to take the standardized survey through internet, mail, telephone, or in-person interviews. These addresses are selected through a sampling method that ensures that a more stable estimate for sparsely populated areas and groups (Groves, 2012). Since the survey is mandatory, the frame and actual sample are very similar. The Census Bureau is bound to strict confidentiality and has employed statistical methodologies so that the data we have access to has no identifying information. Strengths of the ACS would be its large scale response rate and several topics covering housing, social, and economical characteristics. However, the ACS lacks other key variables such as precise household vicinity, political ideology, and religion.

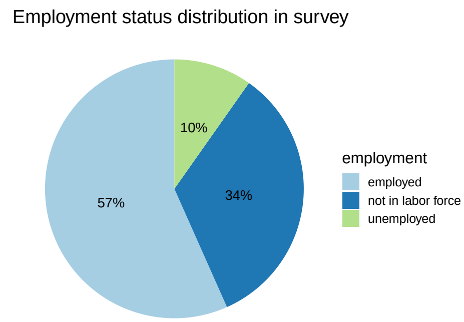
While the ACS data measures hundreds of variables, we extracted 10 demographic variables we thought could predict the popular vote in the 2020 American presidential election. They were: age, sex, household income, education attainment, employment status, state, region, birthplace, race, and whether a respondent had Hispanic origins. These variables were also chosen because of their ability to coincide with that of the survey variables. For example, household income was chosen rather than individual income since the UCLA survey did not ask about personal income. We further manipulated the data by cleaning responses to match the UCLA options. For income, we constructed bins of income intervals that match the survey's since the ACS had exact income values. We selected respondents between 18 and 93 years of age. We made birthplace a binary variable to be either born in the 'USA' or 'another country'. We constructed a new race variable that incorporates Hispanic origins; this meant that if a respondent had answered they had Hispanic origins, it would override and replace the answer of their identifying race. We also kept Chinese identifying respondents separate from other Asians and did these things because Chinese and Hispanic respondents have shown to have strong voting trends with contemporary topics like America's border policies and COVID-19 (Krogstad and Lopez, 2020).

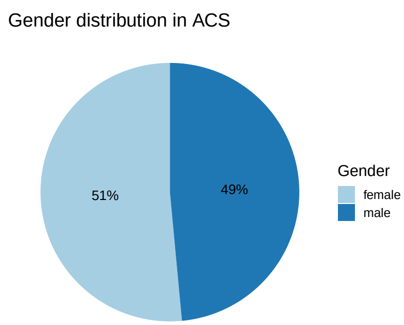
The figures below compare the variables of the ACS data after cleaning with the survey data. From the plots generated from the survey data, it was observed that the majority of respondents were aged between 30 to 44 years, and were split evenly between male and female respondents. The majority of respondents were observed to be employed at 57% of the respondent population, followed by respondents not actively in the work force at 34% (Figure 3). The majority of respondents were White (69%), followed by Native American (14%), and Black Americans (10%), and other races ranged between 1-3% of the respondent population (Figure 4). Based on the survey, a greater distribution of respondents were seen to be more likely to vote for Joe Biden (Democratic candidate) in the upcoming 2020 election compared to Donald Trump (Republican candidate), at 51% and 48% of respondents answering respectively (Figure 5). This may be due to respondent bias, as the Nationscape dataset itself was in partnership with the Democracy Fund, it is expected that a greater number of respondents who support the democratic party would respond.

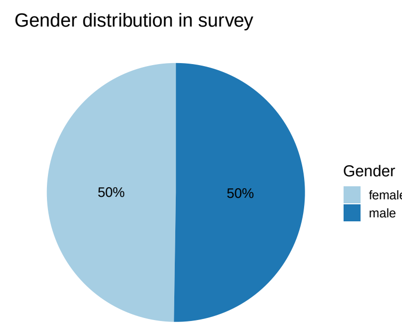

Age distribution in ACS
relative frequencies
0%
10%
20%
18 to 29
45 to 59
30 to 44
60 to 74
74 and above
age

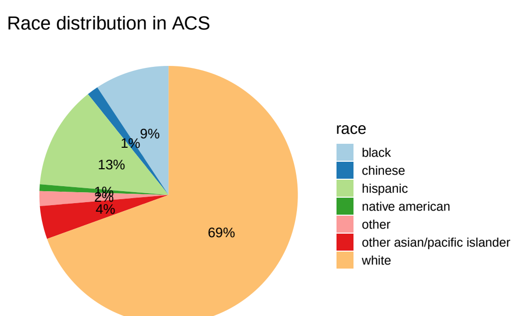

Age distribution in survey
relative frequencies
0%
10%
20%
30%
45 to 59
74 and above
18 to 29
30 to 44
60 to 74
age

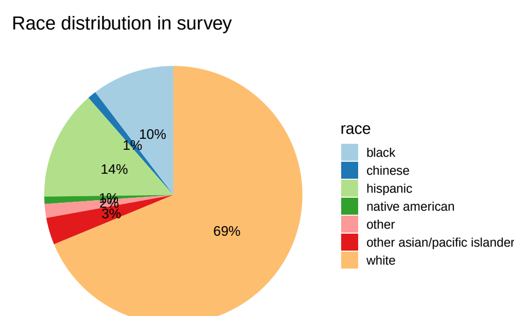

Employment status distribution in ACS
3%
39%
59%
employment
employed
not in labor force
unemployed


Employment status distribution in survey
10%
34%
57%
employment
employed
not in labor force
unemployed


Gender distribution in ACS
51%
49%
Gender
female
male


Gender distribution in survey
50%
50%
Gender
female
male


Race distribution in ACS
9%
1%
13%
4%
69%
race
black
chinese
hispanic
native american
other
other asian/pacific islander
white


Race distribution in survey
10%
1%
14%
3%
69%
race
black
chinese
hispanic
native american
other
other asian/pacific islander
white

3 Model

Before going into the details of the post-stratification we conducted, we will discuss what post-stratification is and when it can be used. Post-stratification is implemented after simpler random samples have been conducted. It is used to properly balance the representation of variables across the target population to gain more precise estimates and create greater confidence in inferences being made. Post-stratification involves classifying each member of a population into a single subgroup or strata to then calculate the probability sample from each stratum. Because of this, post-stratification cannot be used in studies with observations that overlap or are not clearly classified, as this may inaccurately reflect the population. Another challenge is to find definitive lists of variables for an entire population that fall in line with the variables collected for the original sample.

4 Results

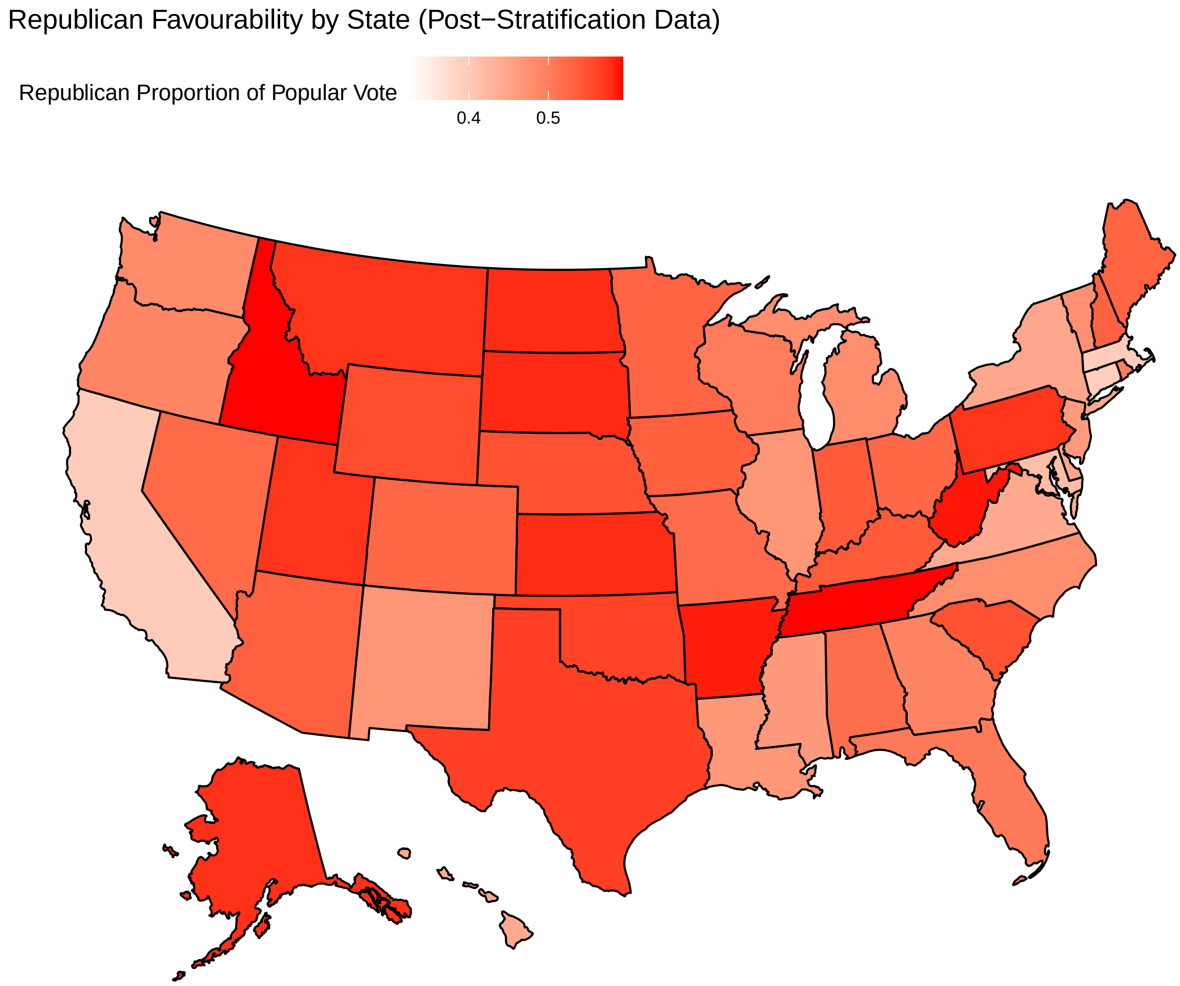


Figure 1: Trump favorability by state.

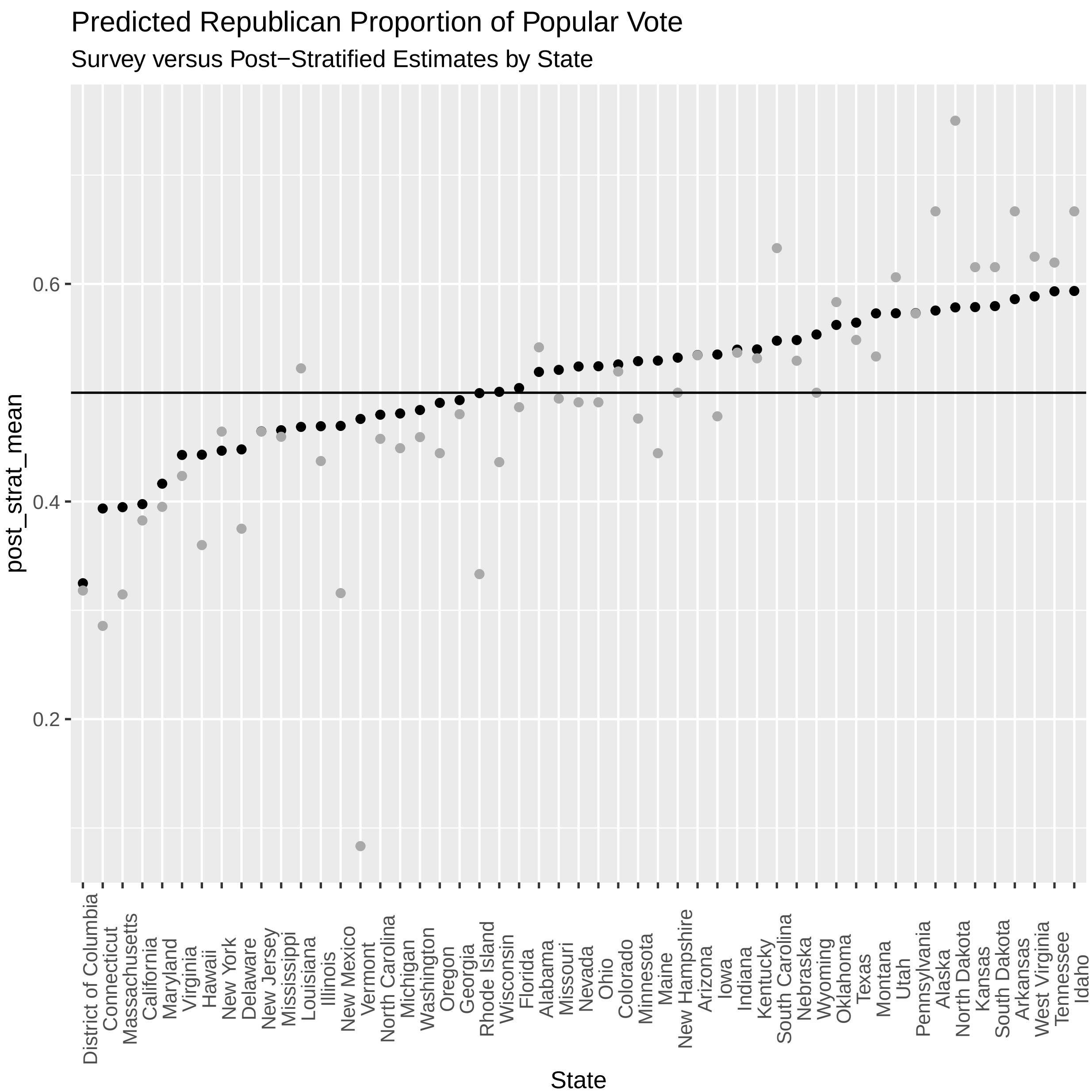


Figure 2: Survey versus poststrat by state.

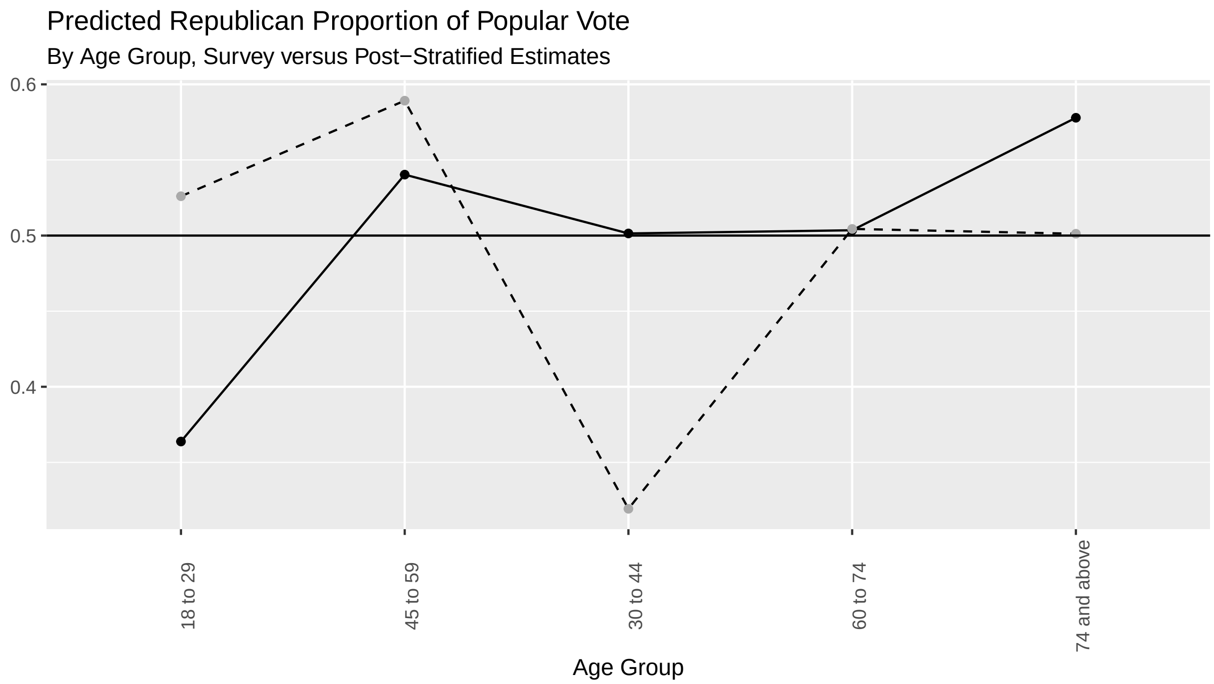


Figure 3: Survey versus poststrat by age

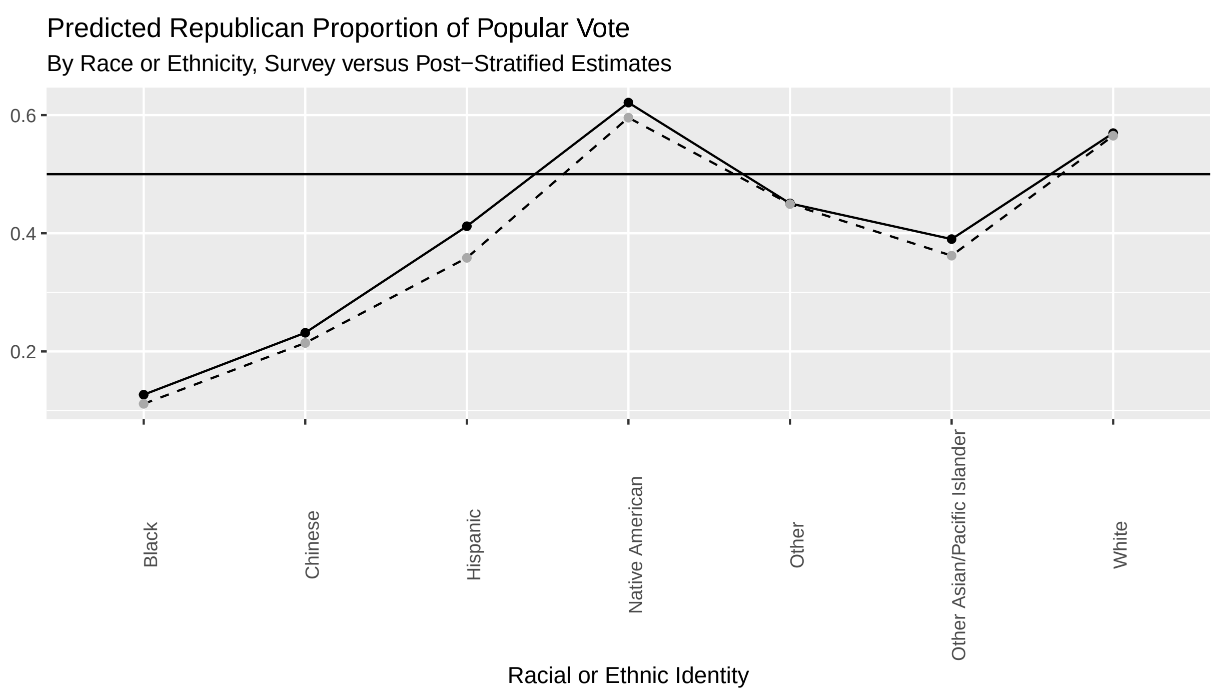


Figure 4: Predicted Republican vote by race

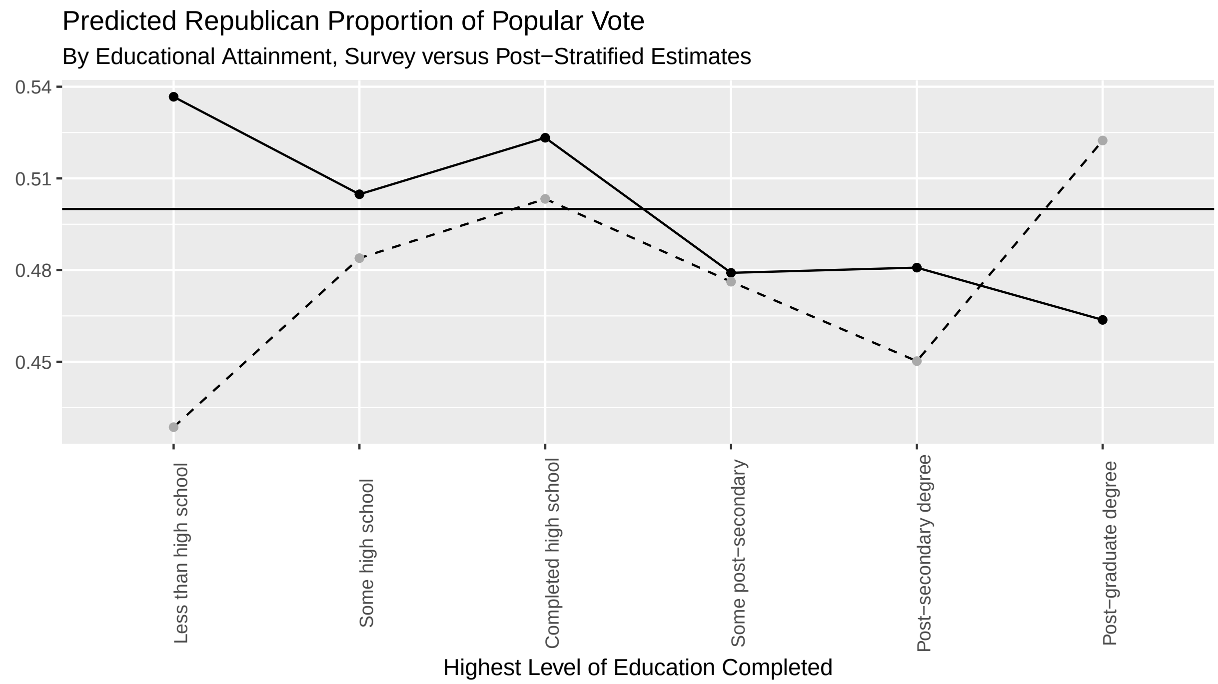


Figure 5: Education post-stratification estimates

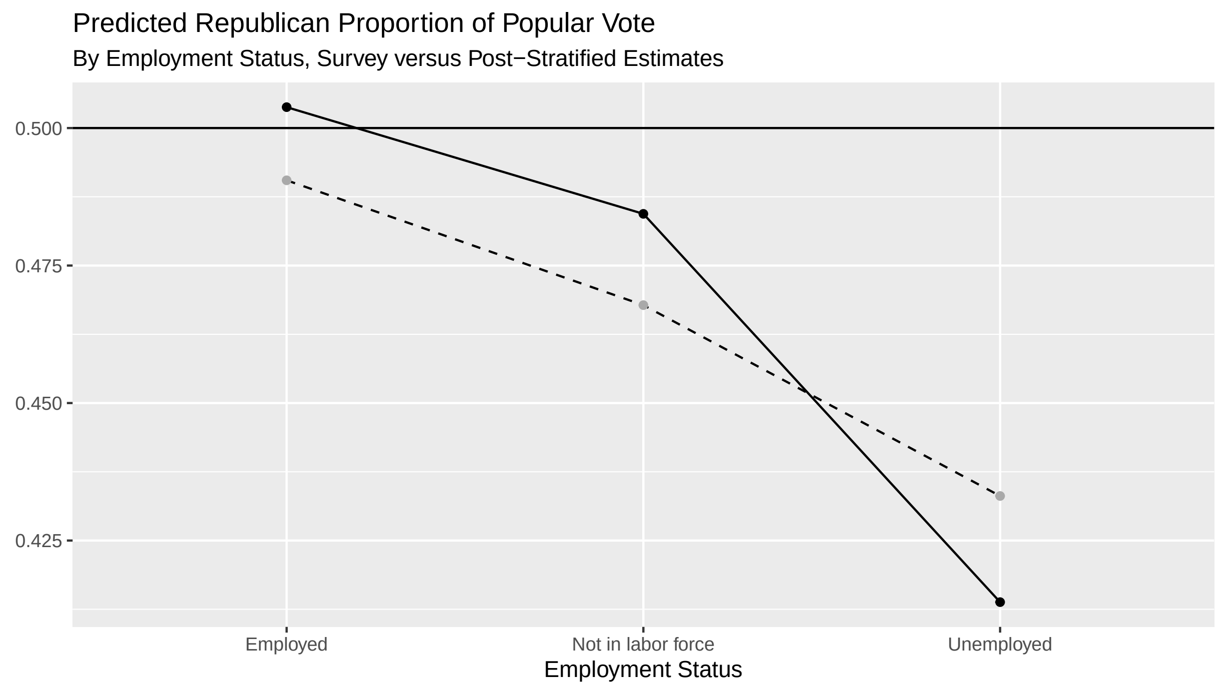


Figure 6: Employment post-stratification estimates

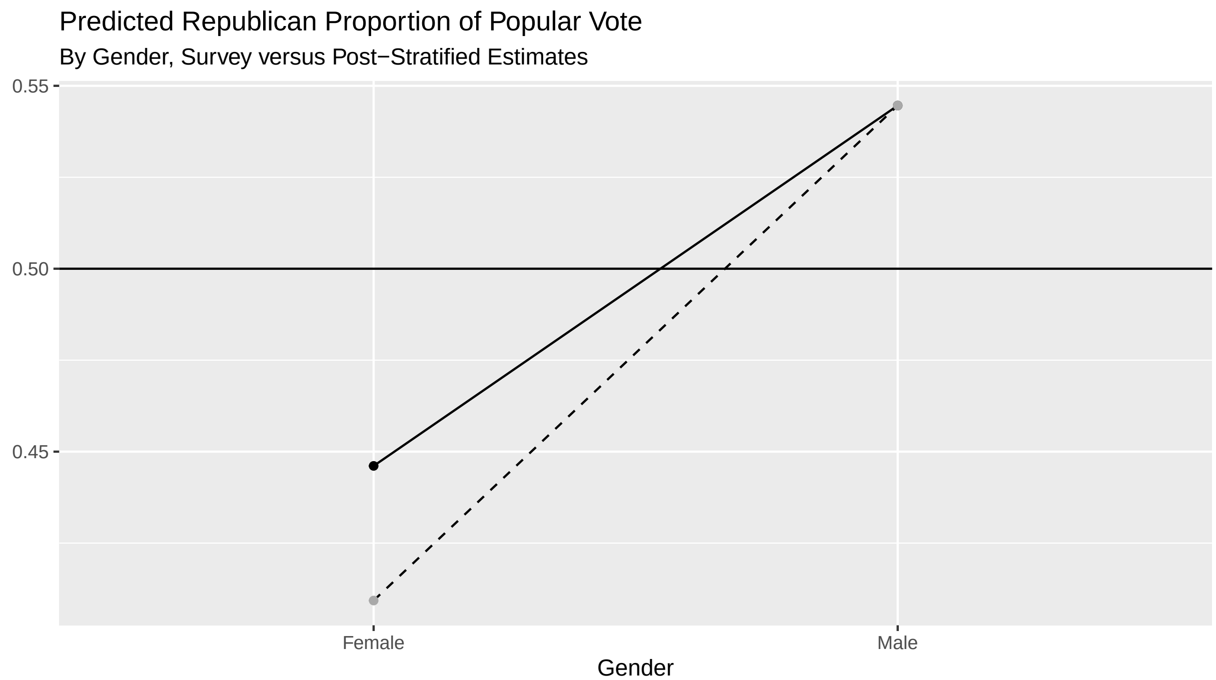


Figure 7: Gender post-stratification estimates

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 signficiant 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

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.

Groves, R. (2012). The pros and cons of making the Census Bureau's American Community Survey Voluntary. *Prepared statement of Robert M. Groves, Director of the U.S. Census Bureau, before the Subcommittee on Health Care, District of Columbia, Census and the National Archives Committee on Oversight and Government Reform United States House of Representatives.* 6 March 2012.

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.

Krogstad, J. & Lopez, M. (2020). Latino voters have growing confidence in Biden on key issues, while confidence in Trump remains low. Pew Research Center. Retrieved from: https://www.pewresearch.org/fact-tank/2020/10/16/latino-voters-have-growing-confidence-in-biden-on-key-issues-while-confidence-in-trump-remains-low/

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.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Silver, N. (2012). *The signal and the noise: why so many predictions fail–but some don't.* Penguin.

Tausanovitch, C. & Vavreck, L. (2020). Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/.

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

Di Lorenzo, Paolo (2020). usmap: US Maps Including Alaska and Hawaii. R package version 0.5.1. https://CRAN.R-project.org/package=usmap

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

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