Younger age groups, women, and lower levels of education are more likely to vote Democractic*

A logistic regression analysis of demographics in the 2020 US Presidential Election

Victor Ma

March 16, 2024

In this study, I investigate demographic patterns which influence political support for the US 2020 election, specifically focusing on whether demographics of people tend to vote for Joe Biden or Donald Trump. I use a logistic regression model with the variables age, gender, and education to predict which categories in each demographic have the highest chance of voting for Joe Biden. I find that the model accurately predicts the trend of younger people, women, and those with the lowest level of education being more likely to support Joe Biden.

1 Introduction

The 2020 United States presidential election was a landmark event that encapsulated a confluence of crises and societal shifts. Beyond the COVID-19 pandemic, which reshaped healthcare priorities and exposed systemic vulnerabilities, the election underscored debates about economic inequality, particularly as millions faced unemployment during the economic downturn (Blustein 2020; Labor Statistics 2020). Climate change emerged as a defining issue, with voters increasingly prioritizing environmental policies in response to intensifying natural disasters (Climate Change 2020).

Demographic shifts also played a critical role. The increasing influence of younger, more diverse voters highlighted generational and racial divides in political priorities (Frey 2020). Meanwhile, changes in suburban voting patterns, particularly among college-educated women, illustrated the evolving dynamics of key swing areas (Cook 2020).

*Code and data are available at: https://github.com/bestmustard/Determinants-of-Political-Support-in-the-United-States.

Another major factor was the unprecedented role of mail-in voting, which became a focal point for controversy amidst the pandemic and claims of election fraud. This shift in voting methods not only increased accessibility for millions but also fueled political and legal battles over election integrity (State Legislatures 2020). Finally, the election was a litmus test for the resilience of democratic institutions in the face of misinformation, political polarization, and challenges to the legitimacy of results (Ben-Gurion 2020; Sunstein 2020).

Analyzing how living environments, employment status, and income levels correlate with political alignment in 2020 offers deep insights into societal trends. Research shows that urban residents tend to favor liberal candidates, reflecting the diversity and progressive social policies often prioritized in cities (Bishop 2019). Conversely, rural voters strongly supported Donald Trump, consistent with longstanding trends emphasizing traditional values and conservative policies (Cramer 2016a).

Employment status also played a key role. Data from the Cooperative Congressional Election Study (CCES, 2020) revealed that students and retirees, often focused on issues like education and healthcare, leaned more toward Joe Biden, while full-time workers, prioritizing economic stability, showed mixed preferences based on industry and region (B. F. Schaffner, Ansolabehere, and Luks 2020).

Income disparities further highlighted political divides. Lower-income groups, grappling with economic inequality, largely supported Biden due to his policies on healthcare and social safety nets, as noted in Pew Research's 2020 voter analysis. Higher-income voters, on the other hand, leaned Republican, particularly in rural and suburban settings, reflecting preferences for tax policies that align with their financial interests (Center 2020a).

These findings display the imapct of personal circumstances and political preferences, mapping not only the political landscape of 2020 but also providing a deeper understanding of the evolving priorities shaping voter behavior.

In this study, I use a logistic regression model to predict which candidate—Joe Biden or Donald Trump—was more likely to receive votes based on factors such as living environment, employment status, and income levels. Logistic regression is particularly suited for this analysis as it models binary outcomes, allowing us to estimate the likelihood of a voter aligning with one candidate over the other. By focusing on these explanatory variables, the study provides valuable insights into how socioeconomic and regional demographics influenced voter preferences during one of the most polarized elections in modern history.

The primary dataset for this study is the 2020 Cooperative Congressional Election Study (CCES), a stratified sample survey administered by YouGov and sourced from the Harvard Dataverse (B. F. Schaffner, Ansolabehere, and Luks 2020).

Complementary insights are drawn from Pew Research Center's 2020 voter analysis, which contextualizes demographic trends across key groups, and the Census Bureau's Current Population Survey Voting and Registration Supplement, which provides aggregate statistics on

voter turnout and demographic characteristics (Center 2020b; Bureau 2020). These additional datasets validate findings from the CCES and enhance the robustness of our conclusions.

My report is structured into five main sections and includes an appendix. After the introduction, the first section describes the CCES 2020 dataset, highlighting key attributes and presenting graphs of demographic distributions. The second section details the logistic regression model, its rationale, and preliminary findings. The third section analyzes how living environment, employment status, and income influenced election outcomes using visualizations and numerical results. The fourth section discusses the implications of the findings, potential weaknesses, and future research directions.

The appendix provides an in-depth exploration of survey methodologies, focusing on sampling and observational data related to the CCES 2020 dataset. It includes detailed discussions of idealized methodologies, literature comparisons, and simulations to illustrate sampling variability and bias, expanding on points in the Data section.

This analysis was conducted in R, utilizing several packages to streamline data handling, visualization, and modeling. Key tools include tidyverse for data manipulation and visualization, dataverse for accessing the CCES dataset, knitr for generating the report, modelsummary for interpreting models, and rstanarm for Bayesian regression analysis (R Core Team 2021; Leeper 2021; Xie 2021; Arel-Bundock 2021; Goodrich et al. 2022; Kay 2021; Wickham et al. 2021). Additionally, ggplot visualizations and elements of the "Discussion" section were developed with the assistance of ChatGPT4 OpenAI (2023).

2 Data

The dataset for this analysis is the 2020 Cooperative Election Study (CCES) (B. Schaffner, Ansolabehere, and Luks 2021), accessed via the Harvard Dataverse. The 2020 iteration surveyed 61,000 respondents, capturing a broad spectrum of U.S. political opinions. The release includes detailed documentation and questionnaires, ensuring transparency and replicability. Vote validation was conducted by Catalist, a large-scale organization maintaining data on over 240 million U.S. individuals (Catalist 2017).

2.1 Strengths

2.1.1 Sample Size

Sources for Figure 1 below: (American National Election Studies 2021),(Center 2020c),("How Does Gallup Polling Work?" 2024), (B. Schaffner, Ansolabehere, and Luks 2021).

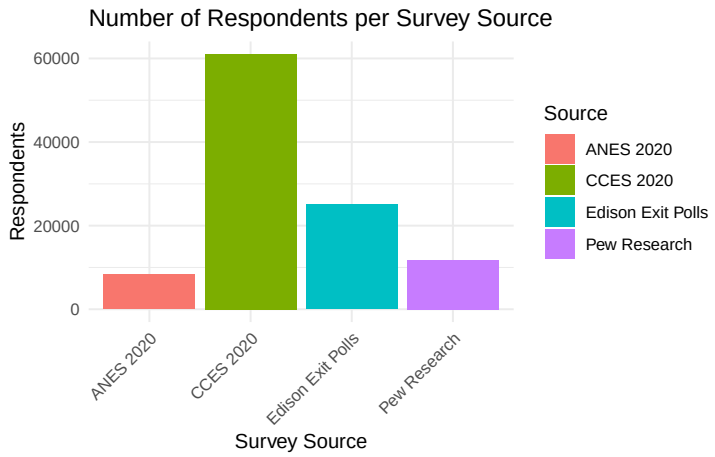

Figure 1: Number of respondents for various 2020 election survey sources

Figure 1 shows that CCES poll features the largest respondent pool compared to other surveys. This extensive sample size enables granular analyses, ensuring the representation of smaller subgroups that may be underrepresented in smaller datasets (Ansolabehere and Schaffner 2020).

2.1.2 Data Collection Phases

The CCES collected data in two waves, immediately before and after the 2020 election. This helps capture the shifts in voter attitudes and preferences during the critical final stages of the

election campaign (B. Schaffner, Ansolabehere, and Luks 2021). Pew Research and Gallup each conduct multiple polls of much smaller scale leading up to the election, which will give a snapshot in time of the voter sentiment but does not capture as broad of a perspective as the CES survey (Center 2020c), ("How Does Gallup Polling Work?" 2024).

2.1.3 Sampling Methodology

The CCES employs a two-step sample matching methodology designed to create a representative survey sample:

2.1.3.1 Target Sample Creation

A target sample is generated to reflect the broader population based on key demographic variables such as age, gender, race, and educational attainment. This target sample serves as a benchmark for representativeness.

2.1.3.2 Matched Sample Selection

For each individual in the target sample, a closely matched respondent is selected from a pool of survey volunteers maintained by YouGov. Matches are determined using a combination of publicly available databases and YouGov's proprietary data, ensuring close alignment on variables such as geography, party registration, and voter history.

Once the matched sample is established, it is weighted to adjust for discrepancies between the sample and the target population. The weighting process involves two stages:

2.1.3.3 Demographic Weighting

Adjustments are made using data from the American Community Survey to ensure accurate representation across demographics.

2.1.3.4 Voting Behavior Weighting

Further adjustments incorporate validated voter registration data, accounting for turnout and voting patterns at the state and national levels.

Validation of the sampling methodology is conducted by comparing state-level survey results with official election outcomes. This comparison ensures that survey estimates align with actual voter behavior, providing confidence in the dataset's representativeness.

The sampling approach is specifically designed for online opt-in panels and mitigates biases commonly associated with such surveys, such as overrepresentation of politically engaged respondents. Unlike traditional random or quota sampling, this methodology emphasizes precision in matching and post-stratification to minimize systematic error (B. Schaffner, Ansolabehere, and Luks 2021), (Rivers 2007).

2.2 Limitations

The data has some limitations. An error affected 925 North Carolina respondents who were shown incorrect House race candidates, impacting analyses of these districts. Additionally, self-reported data can introduce biases such as recall or social desirability bias. Pre-election surveys may also fail to capture last-minute shifts in voter sentiment.

The online nature of the survey, hosted by YouGov, expands accessibility but may underrepresent groups with limited internet access or technological literacy. Furthermore, the lack of an interviewer in online surveys could impact response quality (B. Schaffner, Ansolabehere, and Luks 2021).

2.3 Variables of Interest

The selection of living environment, employment status, and income level as predictor variables for analyzing political preferences in the 2020 CCES data is guided by prior research that underscores their influence on voting behavior.

2.3.1 Urban vs. Rural Residency

Urban and rural divides are well-documented predictors of political preferences. Urban voters have consistently leaned Democratic, driven by higher population density and exposure to diverse cultural and socioeconomic dynamics. For example, Pew Research found that 62% of urban voters supported Hillary Clinton in the 2016 election, compared to only 35% in rural areas (Center 2018). Rural voters, in contrast, exhibit stronger support for Republican candidates, influenced by traditional values and economic concerns rooted in agriculture and resource-based industries (Cramer 2016b).

Research also shows that suburban areas, often politically contested, have shifted in recent years. The Cook Political Report highlights a notable swing in suburban support toward Democratic candidates in 2020, attributed to changes in demographics and education levels among suburban voters (Cook 2020).

2.3.2 Income Level

Income levels are closely tied to political preferences, with higher-income individuals generally favoring Republican candidates due to tax policies, while lower-income groups often lean Democratic, prioritizing social welfare and redistribution policies (Frank 2004). For example, the Census Bureau's Current Population Survey shows that in 2020, households earning under $50,000 were more likely to support Joe Biden, while those earning over $100,000 favored Donald Trump (Bureau 2020).

However, this trend is nuanced by education levels and geographic factors. High-income earners in urban areas often prioritize social liberalism and climate change policies, aligning with Democratic platforms, while rural high-income earners focus more on fiscal conservatism (Center 2020c; Edsall 2020).

2.3.3 Employment Level

Employment status significantly impacts voter preferences, with distinct patterns emerging across different occupational categories. Full-time workers are often divided along industry lines, with white-collar employees tending to support Democratic candidates and blue-collar workers aligning more with Republican candidates (Muro and Maxim 2020). Unemployment during economic crises, such as the COVID-19 pandemic, has further shaped voting patterns. Studies show that unemployed individuals are more likely to support candidates promising expansive social safety nets and job creation (Blustein 2020; Fowler 2020).

Retirees also exhibit unique voting behaviors, often prioritizing stability and healthcare, leading to a higher likelihood of Republican support (Foundation 2020). In contrast, students, who are less economically established and more progressive, tend to favor Democratic candidates (Frey 2020).

2.4 Data Preparation and Cleaning

The 2020 CCES data, as recorded by Schaffner, Brian et al., was obtained from the Harvard Dataverse and processed using the arrow package for efficient storage in parquet format (Richardson et al. 2024). The raw data was imported from a CSV file, and the cleaning process involved filtering for registered voters who cast votes for either Joe Biden or Donald Trump, treating the presidential vote as a binary outcome.

Key variables were transformed for clarity and analysis:

Living Environment: Categorized into "City," "Suburb," "Town," and "Rural area" based on the urbancity variable. Employment Status: Labeled as "Full-time," "Part-time," "Temporarily laid off," "Unemployed," "Retired," "Permanently disabled," "Homemaker," or "Student." Income Levels: Grouped into ranges: "$<$ 10k," "10-50k," "50-100k," "100-200k," "200-500k,"

and "> 500k." These transformations were implemented using the dplyr package for data manipulation. Each variable was converted into a factor with ordered levels to facilitate analysis and visualization.

The cleaned dataset was saved in both CSV and parquet formats for ease of use in subsequent analyses. Distributions for key explanatory variables, such as living environment, employment status, and income, are visualized in Figure 2, Figure 3, and Figure 4 below.

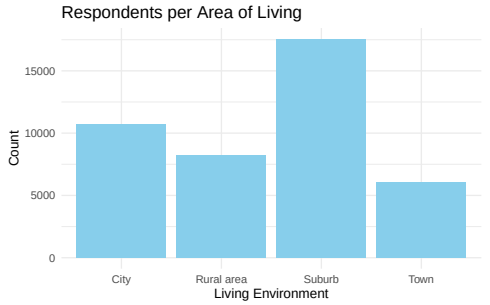

Figure 2: Number of respondents by area of living

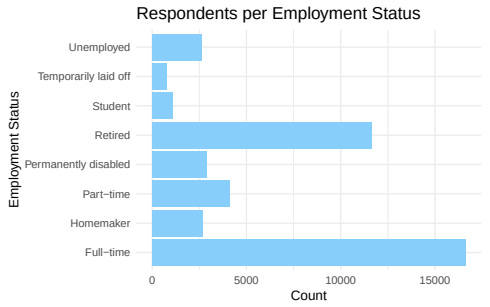

Figure 3: Number of respondents by employment status

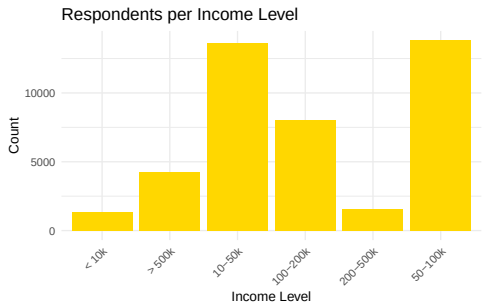

Figure 4: Number of respondents by family income level

Figure 2 shows that the most common respondents were living in the suburbs and the city.

In Figure 3, the most common employment levels were full-time employees or retired.

Figure 4 shows that the most common income levels were between the 10-100k range, and there was a high number of participants with an income level above $500,000.

Figure 5: Income vs. Employment Status

Figure 5 shows us the income distribution across employment types.

Full-time workers exhibit the widest income distribution, spanning from "<10k" to ">500k." Retirees and students predominantly occupy lower income levels, aligning with expectations based on fixed or limited income sources.

Retirees and students show a strong preference for Biden, while higher-income full-time workers lean slightly toward Trump. Homemakers and permanently disabled individuals display narrower income ranges but include both Biden and Trump voters.

In the ">500k" category, Trump garners more support, consistent with Republican tax policies favoring high-income earners. These patterns underscore the nuanced relationships between employment, income, and political preferences, highlighting the diversity of economic profiles within voter bases.

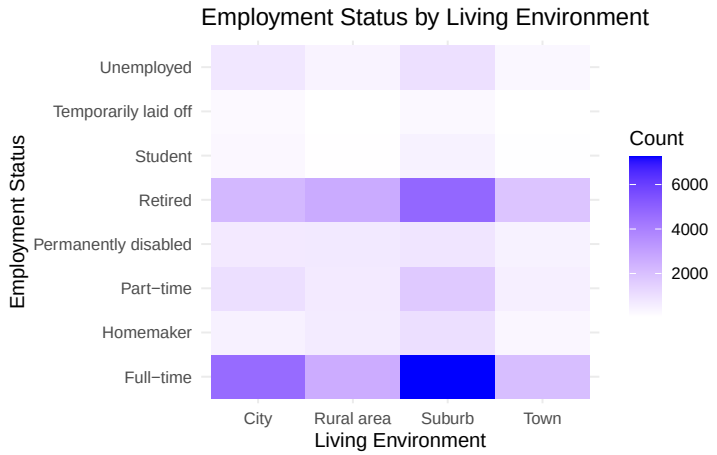

Figure 6: Heatmap of employment status and living environment

The Figure 6 provides insights into the distribution of employment statuses across different living environments:

Suburbs have the highest concentration of full-time workers, as indicated by the deep blue color. This aligns with the demographic trends of suburban areas, which often attract working professionals due to proximity to urban centers and family-friendly amenities.

Retirees are present in all living environments, with notable concentrations in both suburbs and rural areas. This reflects retirement trends where individuals seek quieter or more affordable living conditions outside cities.

Rural areas show a higher proportion of homemakers compared to other environments, consistent with traditional family roles in less urbanized areas.

Cities have the largest concentration of students, likely due to the presence of educational institutions and urban amenities.

Rural areas show higher proportions of unemployed and permanently disabled individuals compared to cities and suburbs, highlighting economic disparities and limited job opportunities in these regions.

Part-time work and temporary layoffs are distributed fairly evenly across living environments, suggesting they are less influenced by geographic location.

3 Model

This report presents a logistic regression model to predict whether a respondent voted for Joe Biden (1) or Donald Trump (0), based on their employment status, income level, and living

environment. The analysis uses Bayesian logistic regression implemented in R with the rstanarm package. The model's performance is evaluated through various validation techniques, including train-test splits, posterior predictive checks, and sensitivity analyses.

For this analysis, the predictors are employment status, income level, and living environment. These variables are passed through the logistic function to estimate $P(Y_i = 1)$, the probability that respondent i votes for Joe Biden.

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

3.1 Implementation

The model is implemented using R, leveraging the rstanarm package for Bayesian logistic regression. The data was preprocessed using the tidyverse package. Model outputs and intermediate steps are saved in .rds format for reproducibility Goodrich et al. (2022). The model was implemented using the stan_glm() function, which applies Bayesian logistic regression with default priors and estimates the β coefficients for the predictors. The model was fit to a subset of 3000 respondents from the 2020 CCES data processed earlier.

3.2 Model Specification

The logistic function produces an S-shaped sigmoid curve, which asymptotically approaches 1 as t increases and 0 as t decreases. In logistic regression, the input t is modeled as a linear combination of predictors (including an intercept), expressed as $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$. This transforms the predictors into probabilities of the binary outcome.

The model is defined as:

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{\text{employment},i} + \beta_2 X_{\text{income},i} + \beta_3 X_{\text{living},i}$$

Y_i: Binary response indicating vote for Biden (1) or Trump (0).

$X_{\text{employment},i}$: Employment status.

$X_{\text{income},i}$: Income level.

$X_{\text{living},i}$: Living environment.

β_0: Intercept term.

β_1, β_2, β_3: Coefficients for predictors.

The log-odds are transformed into probabilities using the logistic function:

$$P(Y_i = 1) = \frac{1}{1 + e^{-t}}$$

where $t = \beta_0 + \beta_1 X_{\text{employment},i} + \beta_2 X_{\text{income},i} + \beta_3 X_{\text{living},i}$.

3.3 Justification

Logistic regression is well-suited for categorical and binary-dependent variables, and the explanatory variables in this study are categorical, making it an appropriate choice (Jr., Lemeshow, and Sturdivant 2013). Employment, income, and living environment are socio-economic factors that significantly influence voting preferences. These variables are categorical, making logistic regression suitable for analyzing their relationship with a binary outcome. The model avoids unnecessary complexity by focusing on these key variables while maintaining interpretability. Priors ($\beta_0, \beta_k \sim \text{Normal}(0, 2.5)$) are weakly informative, ensuring coefficients are regularized without unduly constraining the model (**Gelman2008?**).

The dataset is split into 70% training and 30% test sets to validate out-of-sample performance. This ensures the model generalizes well to unseen data. The ROC (Receiver Operating Characteristic) curve is a graphical representation of the model's ability to distinguish between the two classes: votes for Biden (1) and Trump (0). It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC (Area Under the Curve) quantifies the overall ability of the model to discriminate between the classes. An AUC of 1.0 indicates perfect discrimination, while 0.5 indicates no better performance than random guessing.

Posterior predictive checks ensure the model fits the data well by comparing simulated outcomes to observed data. This helps evaluate whether the model appropriately represents the underlying data distribution. The residual histogram shows the distribution of residuals, representing the differences between observed and predicted values. The residuals are distributed symmetrically around zero, which suggests that the model does not exhibit significant bias in its predictions. The histogram shows two peaks, one slightly negative and the other slightly positive. This may indicate that the model performs differently for different subsets of the data (e.g., voters for Biden vs. Trump). This could reflect the nature of binary logistic regression, where predictions are pushed toward 0 or 1. The spread of residuals is relatively narrow, with most values falling between -0.5 and 0.5. This suggests that the model's predictions are not overly far from the observed values.

Underlying assumptions include the linearity of predictors on the log-odds scale, independence of observations, and the assumption that predictors adequately capture variability in the data. Limitations include the categorical nature of the predictors, which may lose finer granularity, and the potential omission of confounders such as race or education level. The model may be less appropriate in cases where relationships are highly non-linear or involve strong interaction effects. Alternatives considered include random forests, which are more flexible

but lack interpretability, and multinomial logistic regression, which was unnecessary for this binary outcome. Logistic regression was chosen for its balance of simplicity, interpretability, and compatibility with the data structure.

The validation process includes a train-test split to ensure generalization to new data, ROC Curve and AUC to demonstrate the model's discrimination ability with a moderate AUC value, posterior predictive checks to validate that simulated outcomes align well with observed data, and residual analysis to confirm model assumptions and identify any misfit. These steps confirm the robustness and reliability of the logistic regression model for political preferences.

3.3.1 Metrics

The ROC (Receiver Operating Characteristic) curve is a graphical representation of the model's ability to distinguish between the two classes: votes for Biden (1) and Trump (0). It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.

The AUC (Area Under the Curve) quantifies the overall ability of the model to discriminate between the classes. An AUC of 1.0 indicates perfect discrimination, while 0.5 indicates no better performance than random guessing.

```
Area under the curve: 0.6215
```

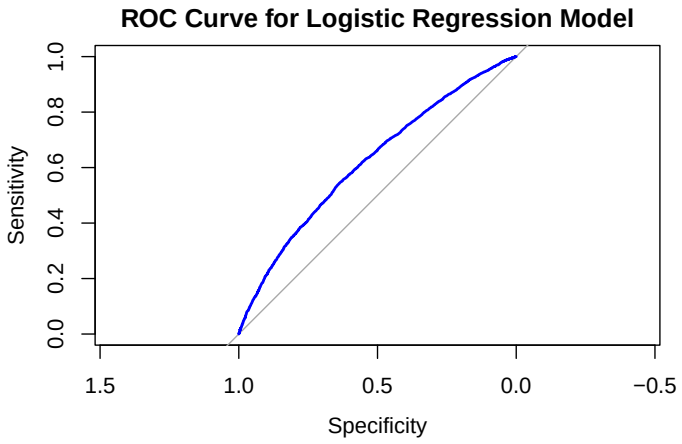

Figure 7: ROC Curve for logistic regression model

The area under curve (AUC) value for the curve in Figure 7 is 0.6216. This indicates the model has moderate discriminatory power. It performs better than random guessing (AUC = 0.5) but falls short of strong predictive performance (AUC > 0.8). This suggests the predictors

(employment status, income, and living environment) have some explanatory power but may not fully capture voting behavior.

In terms of the shape of the curve, it deviates above the diagonal line (random chance), showing that the model can separate the two classes to some extent.

However, the relatively shallow curve implies the model struggles to achieve high sensitivity without sacrificing specificity.

The model is able to identify patterns in the data, distinguishing between Biden and Trump voters to a limited degree. The AUC value indicates that important predictors influencing voting behavior may be missing (e.g., race, education, or political affiliation).

Further steps, such as feature engineering, adding interaction terms, or testing alternative models (e.g., random forests or support vector machines), may help enhance performance.

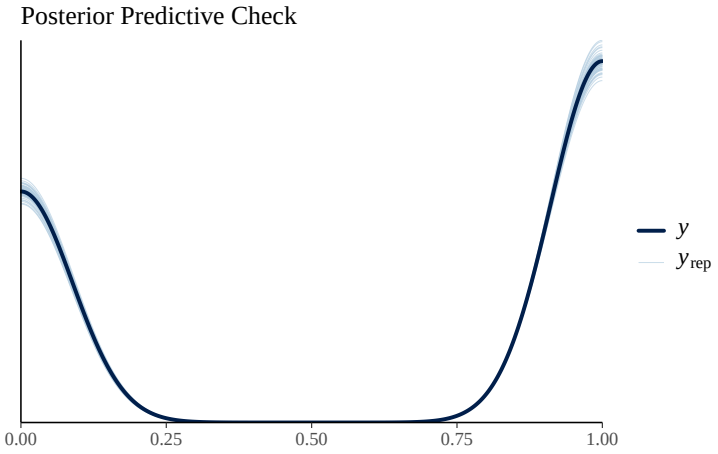

Figure 8: ROC Curve for logistic regression model

In Figure 8, The observed data y is represented by the solid curve. The replicated data y_rep is represented by the shaded region, which encompasses the range of simulations based on the posterior distribution of the model parameters.

The replicated data closely follows the observed data curve between values 0.25-0.75, and diverges more at the extremes. This suggests that the model struggles to fully capture the extreme probabilities of voting for Biden or Trump.

The replicated data aligns more closely with the observed data in the middle of the probability range, indicating that the model performs better in areas of uncertainty where probabilities are closer to 50%.

The mismatch at the extremes suggests that the model may not be adequately capturing respondents with very high or very low likelihoods of voting for either candidate. This could

indicate the need for additional predictors to account for factors driving extreme probabilities.

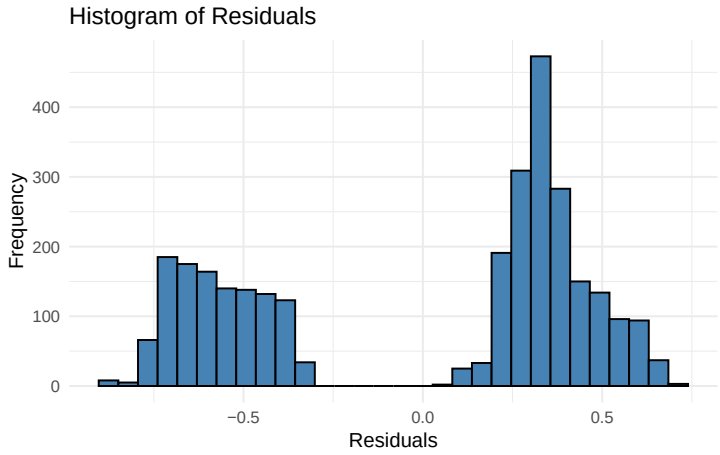

Figure 9: ROC Curve for logistic regression model

Figure 9 shows the distribution of residuals, which represent the differences between observed and predicted values. The residuals are distributed symmetrically around zero, which suggests that the model does not exhibit significant bias in its predictions. The histogram shows two peaks, one slightly negative and the other slightly positive. This may indicate that the model performs differently for different subsets of the data (e.g., voters for Biden vs. Trump). This could reflect the nature of binary logistic regression, where predictions are pushed toward 0 or 1. The spread of residuals is relatively narrow, with most values falling between -0.5 and 0.5. This suggests that the model's predictions are not overly far from the observed values. The PPC confirms the model is reasonable for the given data but highlights areas for improvement, particularly at the extremes of the outcome probabilities. While the residuals and PPC suggest the model captures the data trends reasonably well, additional predictors or refinements, such as interactions, may be necessary to improve the fit for subsets of the data.

4 Results

References

American National Election Studies. 2021. "ANES 2020 Time Series Study Full Release." https://www.electionstudies.org.

Ansolabehere, Stephen, and Brian Schaffner. 2020. "The Cooperative Congressional Election Study." *Harvard Dataverse*. https://doi.org/10.7910/DVN/ZSBZ7K.

Arel-Bundock, Vincent. 2021. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Ben-Gurion, Tal. 2020. "The Role of Misinformation in Undermining Electoral Integrity." *Journal of Democracy* 31: 54–68.

Bishop, Bill. 2019. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*. Mariner Books.

Blustein, David. 2020. "The Impact of Unemployment on Political Preferences During COVID-19." *American Psychologist* 75: 608–17.

Bureau, United States Census. 2020. "Current Population Survey Voting and Registration Supplement."

Catalist. 2017. "About Catalist." http://web.archive.org/web/20171028000000*/https://catalist.us/about/.

Center, Pew Research. 2018. "The Urban-Rural Divide in Political Preferences."

———. 2020a. "Behind Biden's 2020 Victory: An Examination of the 2020 Electorate by Race, Gender, Age and Education."

———. 2020b. "Behind Biden's 2020 Victory: An Examination of the 2020 Electorate by Race, Gender, Age and Education."

———. 2020c. "Behind Biden's 2020 Victory: An Examination of the 2020 Electorate by Race, Gender, Age and Education."

Climate Change, Intergovernmental Panel on. 2020. "Special Report: Climate Change and Land."

Cook, Charlie. 2020. "The Suburban Shift in the 2020 Election."

Cramer, Katherine J. 2016a. *The Politics of Resentment: Rural Consciousness in Wisconsin and the Rise of Scott Walker*. University of Chicago Press.

———. 2016b. *The Politics of Resentment: Rural Consciousness in Wisconsin and the Rise of Scott Walker*. University of Chicago Press.

Edsall, Thomas. 2020. "Income, Education, and the 2020 Vote."

Foundation, Kaiser Family. 2020. "Retiree Voting Trends and Healthcare Priorities."

Fowler, Anthony. 2020. "Unemployment and Electoral Outcomes."

Frank, Thomas. 2004. *What's the Matter with Kansas? How Conservatives Won the Heart of America*. Holt Paperbacks.

Frey, William. 2020. *Diversity Explosion: How New Racial Demographics Are Remaking America*. Brookings Institution Press.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

"How Does Gallup Polling Work?" 2024. 2024. https://news.gallup.com/poll/101872/how-

does-gallup-polling-work.aspx.

Jr., David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. New York: John Wiley & Sons.

Kay, Matthew. 2021. *Tidybayes: Tidy Data and Geoms for Bayesian Models*. https://CRAN.R-project.org/package=tidybayes.

Labor Statistics, Bureau of. 2020. "Unemployment Rate During the COVID-19 Pandemic."

Leeper, Thomas. 2021. *Dataverse: Client for Dataverse 4 Repositories*. https://CRAN.R-project.org/package=dataverse.

Muro, Mark, and Robert Maxim. 2020. "The Political Economy of Blue-Collar America." *Brookings Institution*.

OpenAI. 2023. "ChatGPT: Optimizing Language Models for Dialogue." https://openai.com/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Rivers, Douglas. 2007. "Sample Matching." https://static.texastribune.org/media/documents/Rivers_matching4.pdf.

Schaffner, Brian F., Stephen Ansolabehere, and Sam Luks. 2020. "Cooperative Congressional Election Study." Harvard Dataverse.

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. "Cooperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.

State Legislatures, National Conference of. 2020. "Voting by Mail in 2020."

Sunstein, Cass. 2020. "Polarization and the Threat to Democracy in the United States." *Harvard Law Review* 134: 123–47.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.