My title*

My subtitle if needed

First author Another author

March 11, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Four years ago, the United States once again was at the forefront of an important electoral contest between the Democratic and Republican parties in the 2020 presidential election. The Democratic and Republican parties are represented by Donald Trump and Joe Biden respectively. Donald Trump achieved a remarkable victory in the 2016 election against Hillary Clinton. His winning was surprising and it led people to challenge the preconceived notions and reliability of pre-election polling data [citation needed]. On the other hand, Joe Biden, the former vice-president under Barack Obama's administration from 2008 to 2016, led the Democratic campaign [citation needed]. He aimed to regain control from the Republican party. In the end, Biden won the election by 51.3% majority votes [citation needed]. It is important for us to understand how various demographic, economic, and social factors may influence individuals' voting decisions, which is why we decide to analyze the 2020 election data set and build a prediction model. This allows us to discover trends and address potential issues in society to better reflect the needs and preferences of diverse communities.

In this paper, a logistic regression model is used to forecast the outcome of the 2020 election, with data from the Cooperative Election Study (CES) [citation needed]. Logistic regression is a great choice since it is used to predict binary outcomes, such as election results (Trump or Biden). Our analysis focuses on estimating the likelihood of victory for either Trump or Biden, based on various demographic, geographic, and socioeconomic factors captured in the CES data set, which are race, region that they live, and employment status.

**add 4) what was done; 5) what was found after we finish the results section

*Code and data are available at: https://github.com/hannahyu07/US-Election

This report consists of four sections, not including the introduction. In Section 2, we look at the data used for our report and include some tables and graphs to illustrate the different groups of people in our data. In section 3, we build the model and discuss its justification and explanation. The results of our predictions is highlighted in the Section 4 using tables and graphs. Lastly, discussions are conducted based on the findings, which addresses the voting prediction results based on race, region, and employment status.

In this report, R statistical programming language is used, with R packages: [a lot of citations needed]

1.1 Estimand

2 Data

2.1 Data Measurement

2.2 Summary Statistics

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

```
# read in data
ces2020 <-  read_csv(here::here("data/analysis_data/cleaned_ces2020.csv"))
```

```
Rows: 43534 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (4): voted_for, race, region, employ

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# change column type to factor
ces2020 <-
  ces2020 |>
  mutate(
    voted_for = as_factor(voted_for),
    race = factor(
      race,
      levels = c(
        "White",
        "Black",
        "Hispanic",
        "Asian",
        "Native American",
        "Middle Eastern",
        "Two or more races",
        "Other"
      )
    ),
    region = factor(
      region,
      levels = c(
        "Northeast",
        "Midwest",
        "South",
        "West"
      )
    ),
    employ = factor(
      employ,
      levels = c(
        "Full-time",
        "Part-time",
        "Temporarily laid off",
        "Unemployed",
        "Retired",
        "Permanently disabled",
        "Homemaker",
        "Student",
        "Other"
      )
    )
  ) |>
  select(voted_for, race, region, employ)
```

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in

Appendix B.

3.1 Model set-up

In our analysis, we utilized a Bayesian logistic regression model to examine the relationship between voter preferences and various demographic and socioeconomic factors. The model is formulated as follows:

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) && (1)\\
\text{logit}(\pi_i) &= \alpha + \beta_1 \times \text{race}_i + \beta_2 \times \text{region}_i + \beta_3 \times \text{employ}_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)
\end{aligned}
$$

In this model, y_i represents the binary outcome variable indicating whether an individual voted Biden (as opposed to Trump). The probability of voting for the Biden (π_i) is modeled using a logistic link function (logit(π_i)), which is a linear combination of the intercept (α) and the coefficients (β_1, β_2, β_3)) corresponding to the predictor variables race, region, and employment status, respectively. These predictor variables are denoted as race_i, region_i, and employ_i, where i indexes the individuals in the dataset.

The intercept (α) and coefficients (β_1, β_2, β_3) are assigned informative prior distributions to regularize the model. Specifically, we assume a normal distribution with a mean of 0 and a standard deviation of 2.5 for each parameter.

We chose this modeling approach for several reasons. Firstly, logistic regression is well-suited for binary outcome variables, making it appropriate for analyzing voting behavior. Additionally, Bayesian methods allow us to incorporate prior knowledge and uncertainty into our analysis, providing more robust estimates of the model parameters.

Alternative modeling approaches, such as linear regression models, were also considered. However, we chose Bayesian logistic regression because our result is a binary variable of voter's decision.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`. Rstanarm employs Markov chain Monte Carlo (MCMC) techniques to estimate the posterior distribution of the parameters. To avoid exessive runtime, we randomly sampled 1000 observations to fit the model.Model diagnostics, including convergence checks and posterior summaries, are available in the supplementary materials (see Appendix A).

3.1.1 Model justification

We expect a positive relationship between individuals of Black, Asian, and Hispanic ethnicities and support for Biden. This expectation arises from Trump's history of spreading polarizing language and anti-immigrant sentiments, as well as his controversial plans such as building a border wall. These groups are more likely to align with Biden's policies, which prioritize inclusivity and diversity. White individuals with traditional family values and conservative leanings tend to support Trump. They are drawn to his emphasis on preserving traditional values and promises to uphold conservative principles, especially regarding immigration, law and order, and gun rights.

Conversely, we anticipate a negative relationship between voters in the South and Midwest regions and support for Biden. These regions have a stronger conservative presence and a history of supporting Republican candidates like Trump. States such as Texas and Florida, which are known Republican strongholds, are located in the South. Therefore, individuals in these regions may be less inclined to support Biden's progressive agenda.

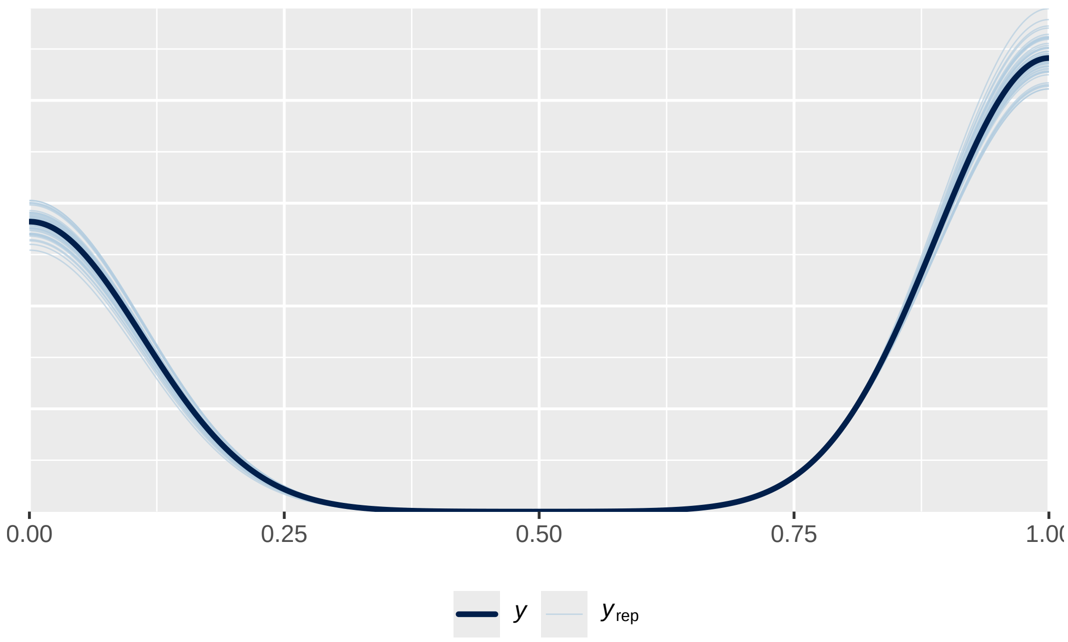
Regarding employment status, we expect students, unemployed individuals, and those temporarily laid off to be more inclined to support Biden. Students are often exposed to diverse perspectives and progressive ideas in educational settings, making them more likely to endorse Biden's platform. Unemployed and laid-off individuals may favor Biden due to the Democratic Party's advocacy for social welfare programs and support for workers' rights.

The voting behavior of employed individuals is harder to distinguish. Some working individuals support Trump due to their opposition to higher taxes and prefer his promises of tax cuts and economic deregulation. Conversely, others lean towards Biden because they believe tax increases should primarily target the wealthy and not burden the middle class. Additionally, educated and liberal-leaning working professionals may prioritize issues such as healthcare, climate change, and social justice, aligning them with Biden's platform.

——————————-needs some explanation

```
# We could use posterior predictive checks, introduced in Section 12.4, to show that the l
political_preferences1 <-
  readRDS(file = here::here("models/political_preferences1.rds"))

pp_check(political_preferences1) +
  theme(legend.position = "bottom")
```



4 Results

4.1 Data Result

Figure 1 shows the relationship between race, region, and voting preference. In the Northeast, Midwest, and West regions of the United States, there is a consistent pattern where individuals across almost all racial groups tend to show greater support for Biden over Trump. Conversely, in the South, there is a distinct divide, with White individuals predominantly favoring Trump, while almost all other racial groups support Biden more than Trump. Overall, the four subplots all illustrate that among Black, Asian, and Hispanic communities, the proportion of support for Biden exceeds that for Trump. However, it is evident that White individuals constitute a significantly larger portion of the population compared to other racial groups, thereby exerting a greater influence on voting outcomes.

Figure 2 illustrates the relationship between region, employment status, and voting preferences for Trump and Biden. Across all regions, people in full-time and part-time employment show a greater percentage of support for Biden than to Trump. Among retired people, support for Trump and Biden is roughly comparable: with a slight tilt towards Biden in the Northeast, Midwest, and West, and towards Trump in the South. In addition, people with other employment statuses generally favor Biden over Trump, except for certain group (Homemaker) in the South.

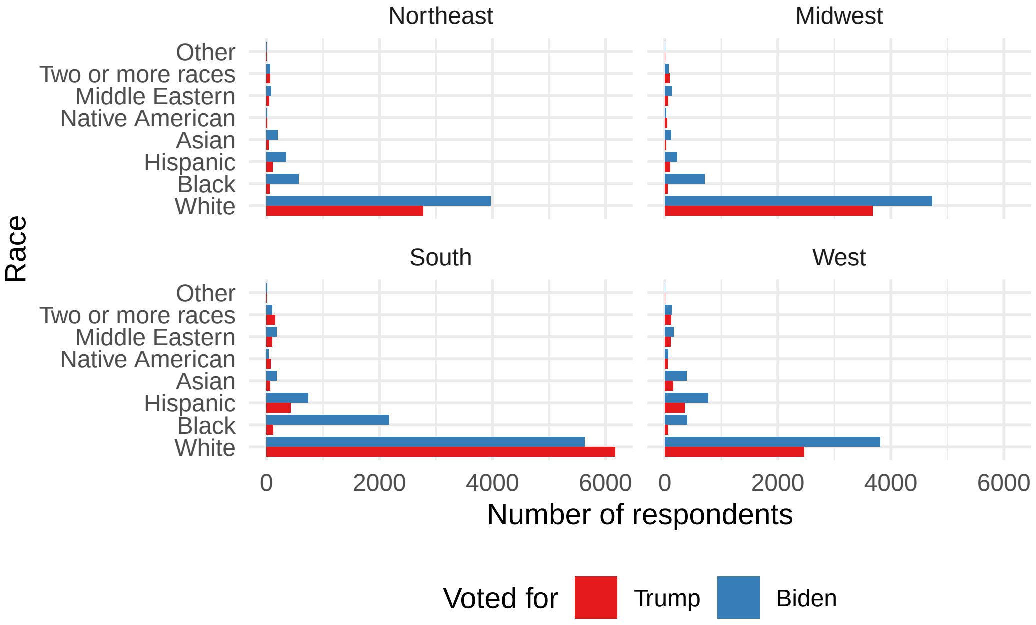


Figure 1: The distribution of presidential preferences, by region and race

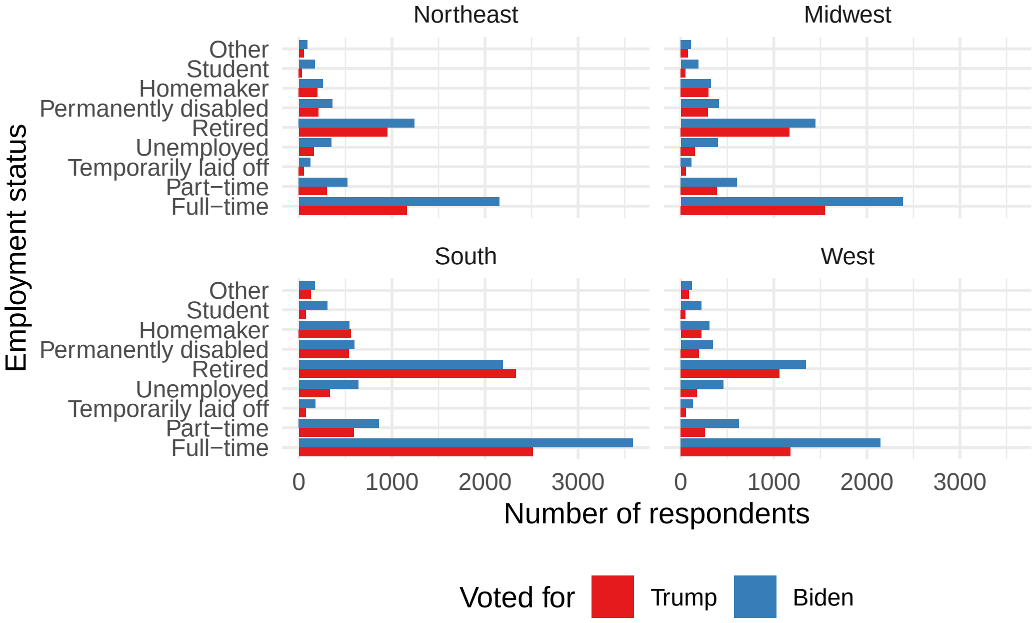


Figure 2: The distribution of presidential preferences, by region and employment status

```
# #| label: fig-race_employ_white
# #| fig-cap: The distribution of presidential preferences, by region and employment statu
# #| echo: false
#
# ces2020_selected_data <- ces2020[ces2020$race == "White", ]
#
# ces2020_selected_data |>
#   ggplot(aes(x = employ, fill = voted_for)) +
#   stat_count(position = "dodge") +
#   facet_wrap(facets = vars(region)) +
#   theme_minimal() +
#   labs(
#     x = "Employment status",
#     y = "Number of respondents",
#     fill = "Voted for"
#   ) +
#   coord_flip() +
#   scale_fill_brewer(palette = "Set1") +
#   theme(legend.position = "bottom")
```

This bubble plot illustrates four variables: race, region, employment status, and voting preference. Darker dots indicate a higher number of people, with red representing Trump supporters and blue for Biden supporters. Purple dots suggest similar support rates for both candidates, since blue and red makes purple in a 1:1 ratio. We can see that the darker dots are mostly in White, Black, Hispanic communities for full-time, part-time, or retired individuals. The darker dots are all blue, which indicates stronger support for Biden. It can also be noted that there are very few entirely red dots, which means that there is no significant disparities in voting percentages between Trump and Biden even among those groups which partially supported Trump. In those purple dots, we found that mostly are from people who are either Native American or from the South.

```
PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installe
```

```
Warning: 'layout' objects don't have these attributes: 'zaxis'
Valid attributes include:
'_deprecated', 'activeshape', 'annotations', 'autosize', 'autotypenumbers', 'calendar', 'cli
```

4.2 Model Results

Our results are summarized in **?@tbl-modelresults**.

Figure 3: The distribution of presidential preferences, by race, region and employment status

```
# # Marginal Effect
# political_preferences_predictions <- predictions(political_preferences1) |>
#   as_tibble()
#
# political_preferences_predictions
```

```
# library(margins)
#
# # Marginal effects for each predictor variable
# marginal_effects <- margins(political_preferences1)
#
# # Plotting marginal effects for race
# race_effects <- dyplot(marginal_effects, variables = "race")
# race_effects
#
# # Plotting marginal effects for region
# region_effects <- dyplot(marginal_effects, variables = "region")
# region_effects
#
# # Plotting marginal effects for employ
# employ_effects <- dyplot(marginal_effects, variables = "employ")
# employ_effects
```

Our results generally matches our justification. To avoid multicollinearity, the model excluded three variables from each category: race Asian, region Midwest, and employed full-time.The intercept represents the estimated log-odds of supporting Biden when all other predictors are held constant at their reference levels. In this case, the estimated log-odds of supporting Biden for individuals who are Asian, live in the Midwest, and are employed full-time is 0.806.

Black individual's support for Biden is large. The estimated coefficient of 2.549 suggests that, holding all other variables constant, Black individuals are estimated to have a 2.549 unit increase in the log-odds of supporting Biden compared to the reference group. Hispanic races on average are also more likely to vote for Biden. We observe an outlier, individuals with race "other" has a 37.999 higher log-odds of supporting Biden compared to the reference group.

Voters preference for candidates also vary by regions and their employment status. Biden has more supports in the Northeast and West region while Trump gains a majority of his votes from the South. As anticipated, students exhibit strong support for Biden. Additionally, individuals who are laid off, unemployed, or employed part-time, potentially reliant on social benefits, also prefer for Biden.

Table 1: Explanatory models Political Preferences (n = 1000)

	Support Biden
(Intercept)	0.806
	(0.438)
raceBlack	2.549
	(0.683)
raceHispanic	0.039
	(0.503)
raceMiddle Eastern	−0.162
	(0.649)
raceNative American	−2.791
	(1.288)
raceOther	37.999
	(33.885)
raceTwo or more races	−1.900
	(0.769)
raceWhite	−0.650
	(0.436)
regionNortheast	0.436
	(0.217)
regionSouth	−0.161
	(0.186)
regionWest	0.137
	(0.206)
employHomemaker	−0.187
	(0.297)
employOther	0.298
	(0.557)
employPart-time	0.433
	(0.270)
employPermanently disabled	−0.273
	(0.295)
employRetired	−0.200
	(0.168)
employStudent	0.831
	(0.501)
employTemporarily laid off	0.699
	(0.553)
employUnemployed	0.416
	(0.331)
Num.Obs.	1000
R2	0.118
Log.Lik.	−606.898
ELPD	−627.2
ELPD s.e.	11.8
LOOIC	1254.4
LOOIC s.e.	23.6
WAIC	1253.7
RMSE	0.46

```
# # Load required libraries
# library(bayestestR)
#
# # Extract posterior draws from the model object
# posterior_draws <- as.matrix(political_preferences1)
#
# # Define the predictor variables for which you want to compute marginal effects
# predictor_variables <- c("Black", "Hispanic", "Middle Eastern", "Native American",
#                          "Other race", "Two or more races", "White", "Northeast",
#                          "South", "West", "Homemaker", "Other employment",
#                          "Part-time", "Permanently disabled", "Retired",
#                          "Student", "Temporarily laid off", "Unemployed")
#
# # Compute marginal effects
# marginal_effects <- marginal_effects(posterior_draws, variables = predictor_variables)
#
# # Summary of marginal effects
# summary(marginal_effects)
```

```
# library(tidybayes)
#
# # Extract draws from the Bayesian model
# draws <- spread_draws(political_preferences1)
#
# # Plot the posterior distribution of a predictor variable
# ggplot(draws, aes(x = raceBlack)) +
#   geom_density(fill = "skyblue", color = "black") +
#   labs(x = "Black", y = "Density", title = "Posterior Distribution of Black Variable")
```

```
# Extract posterior samples
# posterior_samples <- as.matrix(political_preferences)
#
# # Calculate credible intervals (e.g., 95%)
# credible_intervals <- apply(posterior_samples, 2, function(x) quantile(x, c(0.025, 0.975
#
# # Extract coefficient names
# coefficient_names <- colnames(posterior_samples)
#
# # Create a data frame with coefficient names, coefficient estimates, and credible interv
# table_data <- data.frame(
#   Posterior_Mean = colMeans(posterior_samples),
#   Credible_Interval_Lower = credible_intervals[1, ],
#   Credible_Interval_Upper = credible_intervals[2, ]
# )
#
# # Print the table
# print(table_data)
```

Figure # shows range of coefficient estimates of our model within the 90% probability. Due to the fact that the credibility interval for race "Other" is particularly large, we cannot observe the trend of the 90% credibility intervals of other variables. Therefore we created figure # with the x axis limited from -10 to 10.

Combing figure # and figure #, we observe statistical significance for the coefficient estimates for students, the Northeast region, individuals identifying with two or more races, individuals identifying with "Other" racial category, Native Americans, Black individuals, and the intercept, Asians in the Midwest region who are employed full-time. The estimates are significant if the intervals do not cross 0. The value for the estimates are in log-odds, indicating that if the coefficient is positive, the individual supports Biden, if negative, the individual supports Trump.

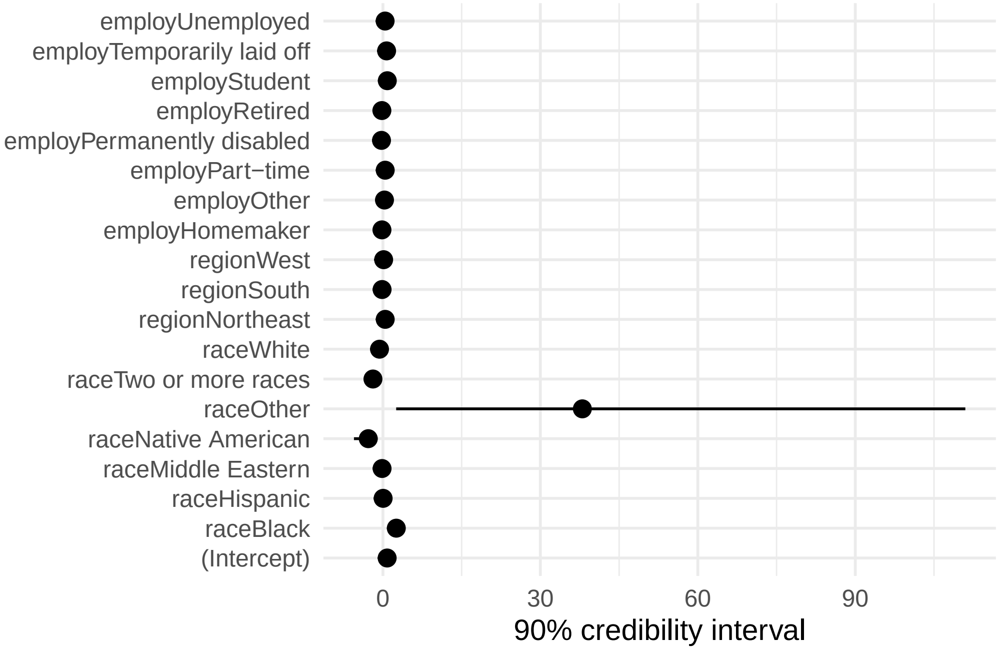


Figure 4: Credible intervals for predictors of support for Biden 1

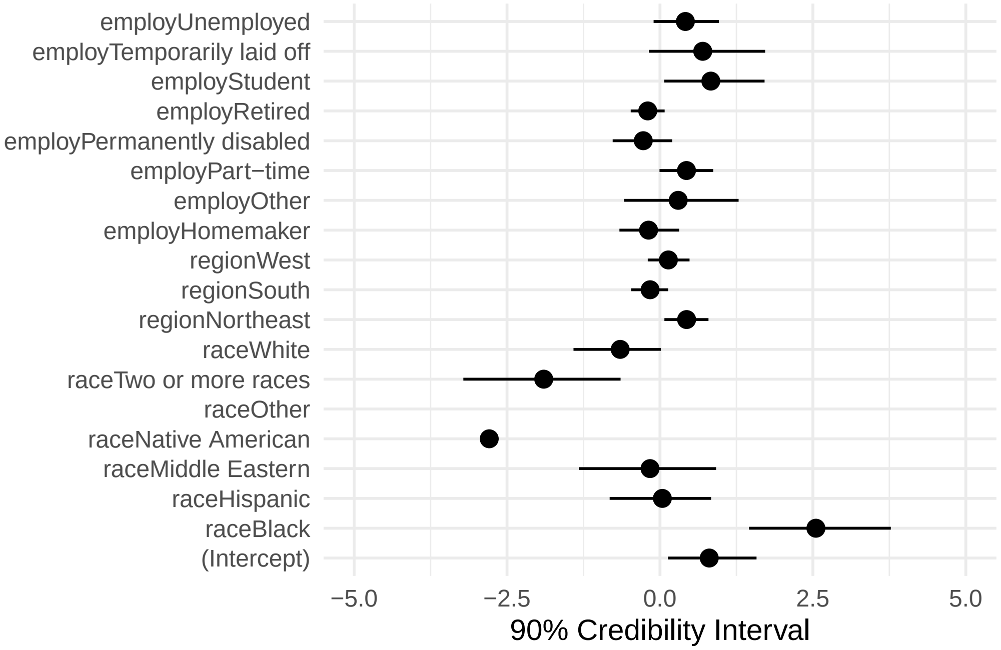


Figure 5: Credible intervals for predictors of support for Biden 2

5 Discussion

5.1 First discussion point

maybe talk about race?

5.2 Regional Variations

Our analysis highlights regional voting variations, with the Northeast, West, and Midwest favoring Biden, while the South supports Trump. These differences stem from diverse cultural, family, religious, and social perspectives.

Biden's support base is concentrated in the Mid-Atlantic states, New England, the West Coast, Southwestern states, and certain regions around the Great Lakes. This support is particularly evident in notable cities within these regions, driven by various factors including liberal ideals and progressive policies. For instance, states like New York and Massachusetts support Biden due to their liberal-leaning populations, diverse demographics, and emphasis on social justices. Similarly, on the West Coast, states such as California and Washington are known for their progressive values and advocacy for issues like environmental conservation and LGBTQ+ rights, contributing to Biden's popularity in these areas. In the Southwest, states like Arizona and Nevada, with growing minority populations and concerns over immigration policies, also show significant support for Biden. Additionally, states around the Great Lakes, including Illinois and Michigan, align with Biden's platform due to their industrial heritage and strong labor unions.

Voters favoring Biden are predominantly concentrated in big cities and large suburban regions, whereas Trump supporters are dispersed across rural areas with lower population density. Consequently, Trump secured victories in 2,588 counties, whereas Biden won over just 551 (Frey 2021). Even in traditionally Republican states, major urban areas exhibit significant support for Biden. Notable examples include Charlotte, North Carolina; Salt Lake City, Utah; and Nashville and Memphis, Tennessee, among others (CNN 2020).

This trend reflects the impact of diverse demographics, progressive ideologies, and evolving societal norms in urban areas. Here, a mix of cultural diversity, acceptance of progressive values, and socio-economic factors favors Biden's platform. Additionally, urban centers, with their concentration of educational and cultural institutions, often align with Democratic policies on healthcare, climate change, and social justice.

The 2020 election witnessed intriguing shifts in traditionally conservative states, turning them into battlegrounds. For instance, Texas, a Republican stronghold, saw unprecedented Democratic gains, though ultimately remaining red. Notably, the five largest cities in Texas—Houston, San Antonio, Dallas, Austin, and Fort Worth—all backed the Democratic party, marking a significant departure from historical voting patterns ("Texas 2020 Election Results" 2020). Similarly, Georgia, a Republican stronghold since 1992, voted for the Democratic party, reflecting evolving demographic and ideological shifts ("Georgia 2020 Election Results" 2020).

Southern states, such as Alabama, Mississippi, and South Carolina with a large amount of Christian population, strongly adhere to conservative principles. For example, these states have implemented strict anti-abortion law and advocated for tight immigration policies. Furthermore, these states have historically been less supportive of LGBTQ+ rights. As a result, voters in these Southern states typically align with the Republican party represented by Trump, which prioritizes pro-life initiatives and traditional Christian values.

5.3 Third discussion point

maybe talk about employment?

5.4 COVID and Mail-in Ballots

Other than race, region, employment, and other usual socioeconomical factors, the special situation of COVID-19 and mail-in-ballots also heavily contributed to the election turnout, leading to Biden's victory. Many Republican's shift from Trump to Biden is attributed to Trump's poor COVID responses and policies.

Mail-in-ballots also spanned many controversy over fraud election as Trump claims

5.5 Weaknesses and next steps

add more weakness about the data, graphs, etc.

One weakness of using a logistic regression model is its inability to predict voting turnout for candidates beyond the Biden or Trump. Expanding the analysis to include other candidates would enhance its comprehensiveness. Furthermore, the reliability of our findings is contingent upon the accuracy of the survey data. Respondents may provide inaccurate or misleading information, intentionally or unintentionally, leading to biased estimates and unreliable predictions. Additionally, missing data (N/As) can introduce further challenges, as logistic regression requires complete data for accurate modeling.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

Figure 6: **?(caption)**

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC
algorithm

Figure 7: **?(caption)**

References

CNN. 2020. “2020 Presidential Election Results.” *CNN*. https://www.cnn.com/election/2020/results/president.

Frey, William. 2021. “Biden-Won Counties Are Home to 67 Million More Americans Than Trump-Won Counties.” *Brookings*. https://www.brookings.edu/articles/a-demographic-contrast-biden-won-551-counties-home-to-67-million-more-americans-than-trumps-2588-counties/.

“Georgia 2020 Election Results.” 2020. *CNN*. https://www.cnn.com/election/2020/results/state/georgia.

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

“Texas 2020 Election Results.” 2020. *CNN*. https://www.cnn.com/election/2020/results/state/texas.