Generational Dynamics, Gender and Politics in the US*

Predicting 2020 US Presidential Voting Preferences

Hritik Shukla

March 16, 2024

This study investigates voting patterns in the 2020 US presidential election, analyzing how political preferences are influenced by a person's gender and the generation they're born in. With the help of exploratory data analysis, logistic regression and data from the 2020 CES Common Consent dataset, it is observed that younger generations tend to favor Biden over Trump. As generations get older, Biden loses more and more support. It is also found that women tend to favor Biden more heavily than the men of the same generation as them.

1 Introduction

'If you're not a liberal when you're 25, you have no heart. If you're not a conservative by the time you're 35, you have no brain.'

Although commonly, but incorrectly attributed to Winston Churchill, this quote summarizes a common belief held in our society - a person generally starts out in life with a left leaning political compass, and as they get older, their beliefs transform into more right leaning ideologies. It is also a popular belief that newer, younger generations tend to be more inclusive, open and liberal in their ideologies (Itkowitz (2019)), and this only intensifies as further generations get old enough to participate in the political landscape of the US.

This paper aims to find out whether these generalizations hold true by observing the 2020 US Presidential Elections. It is important to note that 2020 was a tumultuous year for the entire world, it first year of COVID-19 when not much was known about the disease. President Trump, who was in power before the elections, was greatly criticized for his poor handling of pandemic and lockdown, which lead to massive change in sentimenets amongst the citizens of the country (Jurkowitz (2020)). As a result, with highly charged sentiments, this election

*Code and data are available at: https://github.com/hritikshuklas/vote-preferences-generations saw the highest voter turnout of any presidential election held in the US in the 21st century (*2020 Presidential Election Voting and Registration Tables Now Available* (2021)). Another interesting factor to consider is that this was the first election where Generation Z could participate. Furthermore, they held highest voter turnout of any generation (Hess (2020)). Therefore, it would be even more interesting to observe if these long accepted norms held up in these extraordinary conditions, under situations which would not be present under any other normal circumstances - the first election of its kind in recent memory.

The remainder of this paper is structured as follows:

- Section 2 explores the dataset and the variables within it used for the study
- Section 3 explains our models setup and our assumptions going into the study
- Section 4 explores our findings in detail
- Section 5 evaluates these findings in the context of our reality

I used R (R Core Team (2023)), along with multiple packages to aid in the data analysis and modelling. More particularly, the packages tidyverse (Wickham et al. (2019)), dplyr (Wickham et al. (2023)), tidyr (Wickham, Vaughan, and Girlich (2023)), dataverse (Kuriwaki, Beasley, and Leeper (2023)) and arrow (Richardson et al. (2024)) were used for data acquisition, testing and cleaning. The package rstanarm (Goodrich et al. (2024)) was used for modelling, and ggplot2 (Wickham (2016)), knitr (Xie (2015)) and modelsummary (Arel-Bundock (2022)) were used for data visualization.

2 Data

The final release of the 2020 CES Common Consent dataset by (Schaffner, Ansolabehere, and Luks (2021a)) was used for this study, acquired through the dataverse package (Kuriwaki, Beasley, and Leeper (2023)). This dataset was created from a survey conducted by YouGov, an internet-based data analytics firm based in the UK. Random sample methodology was used to survey 61,000 adults over the internet from September to October 2020 (Schaffner, Ansolabehere, and Luks (2021b)). For our model, the "gender" (Section 2.2) and "CC20_410" (Section 2.3) variables were used from this dataset directly, and the "birthyr" variable was used to construct the "generation" variable (Section 2.4).

2.1 Data Cleaning

The "votereg", "CC20_410", "gender" and "birthyr" variables from CCES 2020 dataset (as seen in Table 1) were selected initially.

As the first step, any voters who weren't registered to vote were removed from our data. The "votereg" variable was used for this process, and then was subsequently removed from our

Table 1: Raw Dataset Preview

votereg	CC20_410	gender	birthyr
1	2	1	1966
2	NA	2	1955
1	1	2	1946
1	1	2	1962
1	4	1	1967
1	2	1	1961
2	NA	1	1950
1	2	2	1947
1	2	2	1970
1	1	2	1963

finalized data as it served no purpose for our model. Next, "CC20_410", which records who the respondent voted for for President of the Untied States, was used to filter out respondents who voted for Joe Biden or Donald J. Trump, as we are interested in peoples' preference towards the Democrats or the Republicans. Any respondents who were born outside the years 1928 to 2012 were removed from the survey, as there wouldn't be enough data for those generations to make any statistical inferences from them.

Next, any responses which were missing any of the these variables we needed were removed during cleaning. This omission culled down our total responses from 61,000 to 43,540. However, we still have more than enough data to derive meaningful results.

Some variables were renamed to be more human-readable - "CC20_410" was renamed to "voted_for" and "birthyr" was renamed to "birthyear". The values for categorical variables (such as "gender" and "voted_for") within the dataset were cleaned to represent their more meaningful, intended values. The questions corresponding to these variables in the survey have a select number of options for the subject to choose from. These responses are recorded in the dataset as a number, where each number corresponds to a relevant response - for example, for the gender variable, a response of 1 means female and 2 means male. During cleaning, these numbered responses were replaced by their true values.

The end result of the cleaning process leaves the dataset shown in Table 2.

2.2 Gender

The gender variable stores the responses of the respondent to a question which asks them to choose between "Male" and "Female" as their only two options to describe their gender. This question would have been better phrased if it were asking for biological sex assigned at birth rather than gender based on the options provided to the respondents. Even though the survey

Table 2: Cleaned Dataset preview

Gender	Birthyear	Generation	Voted For
Male	1966	Generation X	Trump
Female	1946	Baby Boomer	Biden
Female	1962	Baby Boomer	Biden
Male	1961	Baby Boomer	Trump
Female	1947	Baby Boomer	Trump
Female	1970	Generation X	Trump
Female	1963	Baby Boomer	Biden
Female	1966	Generation X	Biden
Female	1961	Baby Boomer	Biden
Male	1959	Baby Boomer	Biden

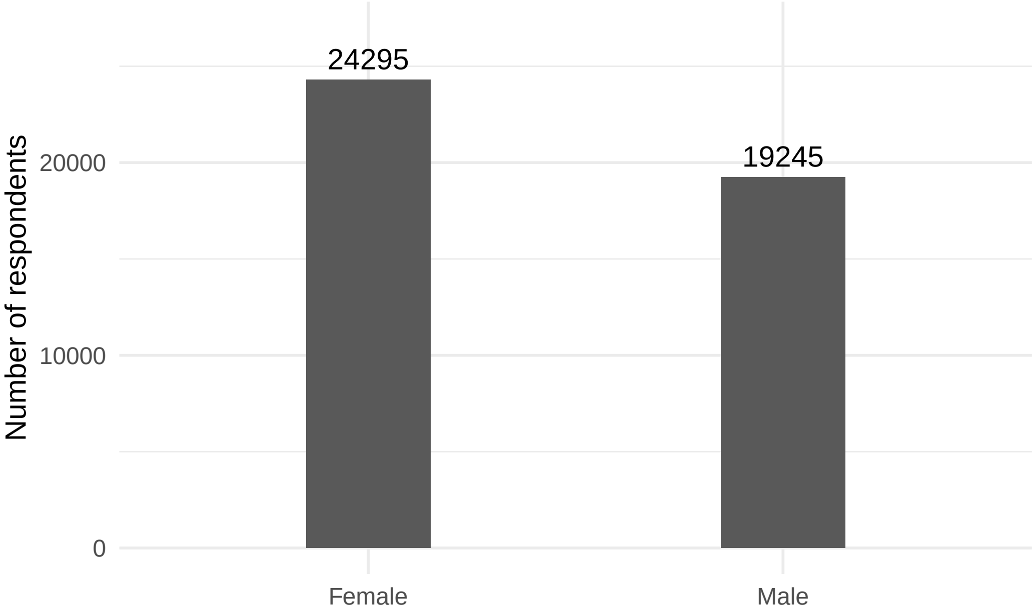


Figure 1: Gender Distribution

asks for the respondent's sexual orientation later, the phrasing of this question alone could be a reason for people to not submit their survey responses as they don't identify with the given options which may lead to under-representation of people belonging to this demographic in the dataset.

In Figure 1, we can see the number of respondents who identified as male and female. According to "The Gender Ratio of United States of America (2020 - 2028, Males Per 100 Females)" (2024), the United States had a gender ratio of 97.14 males to 100 females in 2020. Our data consists of 24,295 females and 19,245 male - the gender ratio represented here is considerably below the national average at the time of the survey - this might introduce some bias into our data.

2.3 Voted For (CC20_410)

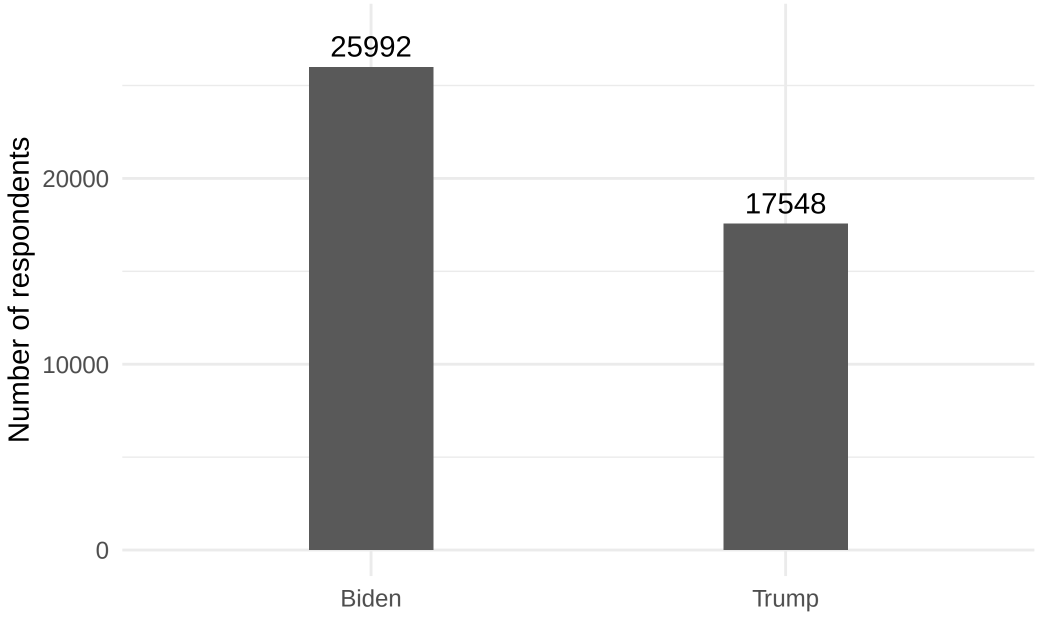


Figure 2: Vote Distribution

The CC20_410 variable, renamed to "voted_for", records who the respondent voted for in the 2020 Presidential Elections. The respondents were given options other than Joe Biden and Donald Trump for this question, such as "Other", "I dd not vote", "Not Sure", etc, but these options were removed as they weren't needed for our purposes.

In Figure 2, we can see that Biden has an overwhelming majority over Trump in terms of popularity, with around a 6,000 vote difference between the two.

2.4 Generation (and birthyear)

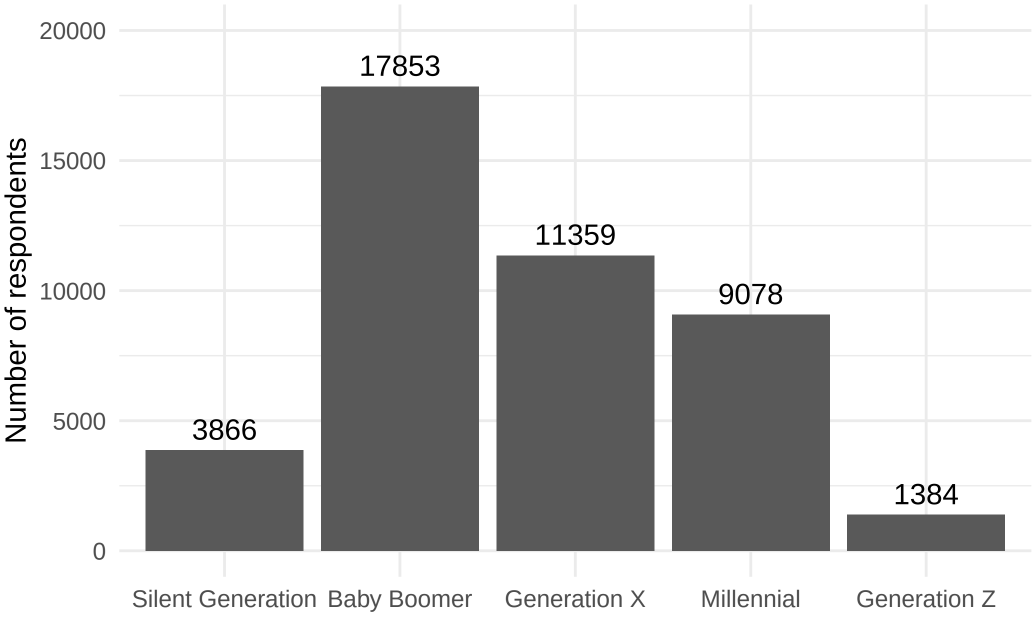


Figure 3: Generation Distribution

The generation variable was created from the "birthyr" variable in the dataset, where respondents recorded their year or birth. Each respondent was assigned a "generation" according to their year of birth. The generation breakpoints used for this paper are defined by Pew Research Center (Dimock (2019)) as follows:

- Silent Generation: 1928 - 1945
- Baby Boomer: 1946 - 1964
- Generation X: 1965 - 1980
- Millennial: 1981 - 1996
- Generation Z: 1997 - 2012

Any respondents whose year of birth falls outside of these year ranges were removed from the dataset. These numbers fall in line with the voter turnout

In Figure 3, we can see that Baby Boomers show the largest representation of all groups. Generation X and Millennial have similar number of respondents, and Silent Generation and Generation Z have the lowest number of respondents. This distribution falls in line with the voter turnout by age statistic reported by Pew Research Center (Gramlich (2020)). Note that even though Generation Z had the highest voter turnout for all generations (Dimock (2019)) and their numbers look low compared to others, it falls in line with expected numbers as the oldest Generation Z respondent who would've been able to vote would've been born in 2002, leaving a considerable population of this generation unable to vote.

3 Model

We will be using logistic regression to model our data, where our outcome variable would be whether a respondent prefers Biden as the presidential candidate. Gender and generation will be used as predictors our outcome variable.

3.1 Model set-up

Define:

- y_i is the political preference of the respondent and equal to 1 if Biden, and 0 if Trump
- gender_i is the gender of the respondent
- generation_i is the generation of the respondent

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) & (1)\\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{generation}_i & (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) & (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) & (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) & (5)
\end{aligned}
$$

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`.

3.1.1 Model justification

Logistic regression is employed for this model since our variable of interest can be constructed as a binary outcome variable - respondent prefers Biden, respondent doesn't prefer Biden (prefers Trump). Logistic regression is well suited for situations where the outcome variable represents two mutually exclusive categories and its probability is based on a set of predictor variables (here, gender and generation).

According to the common beliefs in our society as established in Section 1, we expect to see a positive relationship in the younger generations, who we expect to be more left-leaning and therefore be more favorable towards Biden. And vice versa, we expect to see a negative relationship in the older generations, who we expect to be more right leaning, and therefore, more conservative. However, due to the large dissatisfaction in the majority of US citizens in 2020 due to how President Trump handled the COVID pandemic (Jurkowitz (2020)), this skew might not be as great as it could be for other elections. Regardless, we expect to see a linear relationship between a person's generation and preference for Biden - younger generational cohorts leaning more towards voting for Biden.

The relationship between gender and political preference is a bit more complicated. According to articles from the Conversation (Rosie Campbell (2023)) and the Gallup (Saad (2024)), women tend to vote more conservative than men before 2017, after which this trend flipped on its head and women are began to vote more liberally.

4 Results

Our results are summarized in.

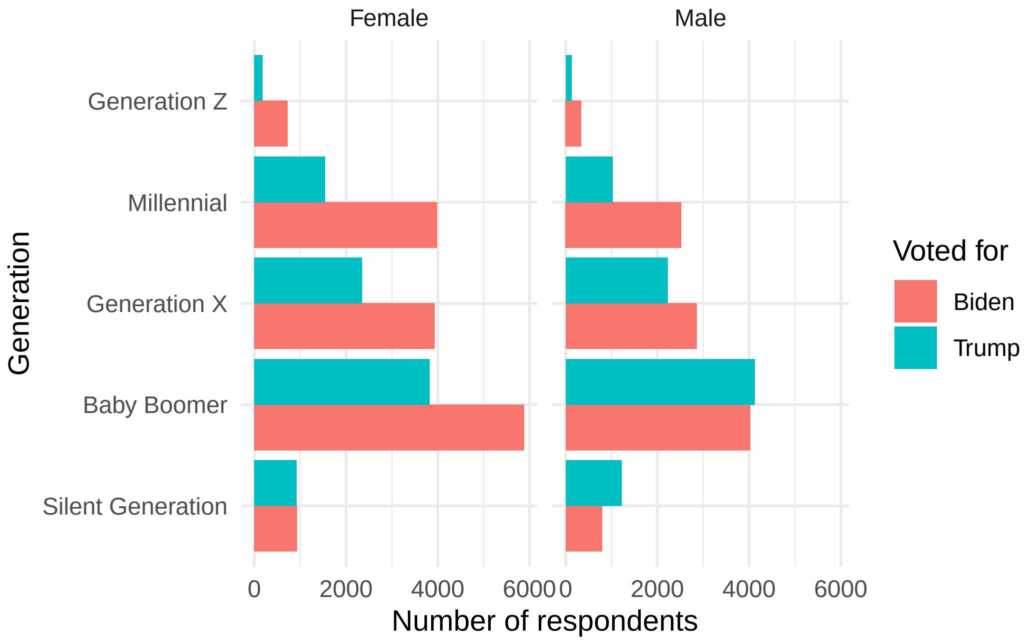


Figure 4: Distribution of presidential preferences, by gender, and by generation

Figure 4 Compares the political preferences between respondents of different generations, separated by gender. The voting behaviour on this histogram closely resembles most of our

Table 3: Model Summary

	Support Biden
(Intercept)	0.051
	(0.035)
generationBaby Boomer	−0.422
	(0.036)
generationGeneration X	−0.597
	(0.040)
generationMillennial	−1.116
	(0.041)
generationGeneration Z	−1.393
	(0.070)
genderMale	0.328
	(0.020)
Num.Obs.	43 540
R2	0.034
Log.Lik.	−28 594.228
ELPD	−28 600.2
ELPD s.e.	54.6
LOOIC	57 200.4
LOOIC s.e.	109.3
WAIC	57 200.3
RMSE	0.48

assumptions, but there are some interesting observations which weren't predicted.

First big observation is the fact that as generational cohorts get younger, the difference in votes cast to Biden and Trump increase in the favor of Biden across both genders, showing that younger voters tend to be more liberally aligned politically. This phenomenon is exaggerated in the female voter base of each generation - the difference between the votes cast to Biden and Trump are much greater compared to those in males.

Between the two genders, female voters showed a much higher discrepancy in their voting preferences in the favour of Biden. This falls in line with the recent trends of women voting more liberal in the recent years.

Note that these observations are further supported by the coefficients of our fitted logistic model, as seen in Table 3:

Note that for our model, we set Biden as our reference level for voted_for, Silent Generation as our reference level for generation, and Female as our reference level for gender. This means that the coefficients of other generations are relative to the Silent Generation, which we have seen from the histogram tend to favor Trump over Biden. Moreover, R treats “Biden” as our failure case and “Trump” as our success case, which implies that negative coefficients imply a positive increase in odds for Biden. Conversely, a positive coefficient imply a decrease in odds for voting Biden, i.e., an increase in odds of voting for Trump. Keeping these key pieces of information in mind, we proceed to our model’s results.

Our model arrived at the intercept of 0.051 with a standard error of 0.035. This means that we can say with fairly high accuracy, that when all predictors are set to 0, i.e, for the female demographic of the Silent Generation, Trump’s log odds for getting a vote is 0.051, which is about 51% probability - this falls in line with our observations in Figure 4.

Keeping our reference level of Silent Generation and Biden in mind, we now observe the coefficients for different generations:

- Baby Boomers have a coefficient of -0.422 with a standard error of 0.036
- Generation X has a coefficient of -0.597 with a standard error of 0.040
- Millennials have a coefficient of -1.116 with a standard error of 0.041
- Generation Z have a coefficient of -1.393 with a standard error of 0.070

Notice that the coefficients tend to decrease further and further as the generational cohorts become younger, implying that Biden tends to gain more favor as the generational cohorts get younger - keeping in line with our hypothesis. Moreover, the standard error for each of these generations is very small, implying a high precision to these predictions made by our model.

Finally, we observe the coefficient for male gender to be 0.328 with a standard error of 0.020, which when compared to the female gender favors Trump over Biden. This implies that male respondents showed an increase of 0.328 log odds in voting for Trump when compared to female respondents. And due to our very low standard error, we can say with high confidence that male respondents have a higher chance of voting for Trump compared to female respondents, which also agrees with our hypothesis.

5 Discussion

5.1 What was done

It is an age old belief that as a person starts out with a left leaning political compass and as they get older, their ideology slowly shifts over to the right. The purpose of this paper was to test this belief by gaining a better understanding of the voting preferences (between Joe Biden and Donald J. Trump) different generations had in the 2020 US Presidential Elections, based on the generation they belonged to, as well as the gender of the members of these generations. This was done by obtaining data from the 2020 CES Common Consent dataset, which contained survey responses from over 61,000 adults surveyed over the internet. Initial exploratory data analysis suggested that younger generations tended to favor Biden over Trump, and women tended to favor Biden more heavily than men of the same generation. Logistic regression modelling was then used to formalize this discovery, and it showed a clear trend which agreed with our initial hypothesis - as generational cohorts get younger, the odds of favoring Biden increase.

5.2 "Younger Generations vote Liberally; Older Generations vote Conservative"

Our main hypothesis of the study was proven to be true from our findings - younger generations vote liberally; older generations vote conservative. From both exploratory analysis and our model, it was found that younger generations, such as Millennial and Generation Z, tend to favor supporting liberal candidates (Biden in this case). As generational cohorts got older, their liberal support slowly decreased until eventually the generation as a whole tended to favor the conservative candidate (Trump). Do note that even though Trump's favorability increased as generations got older, Biden still held majority votes across all generations other than the Silent Generation, who are now a minority in the American voter base. This might be explained by Trump's growing unpopularity in 2020 from the way he handled the COVID-19 pandemic, and over his term as the President as a whole. What's more interesting is that even though only a part of Generation Z participated in the election, and it was the first election for the generation as a whole, they fit in the trend perfectly, showing overwhelming support for Biden in their votes compared to other generations, even more so than Millennials.

5.3 "Women Vote more Liberally than Men"

A number of studies published in the last few years (Saad (2024), Rosie Campbell (2023)) suggested that women were voting more liberally than before, and this was observed in our own findings as well. Against the traditional beliefs that women tended to vote more conservative than men, it was found that women were tended to favor the liberal candidate much greater than the conservative candidates across all generations.

5.4 Weaknesses and next steps

Some weaknesses of this model arrive from the data that was used for this study - the 2020 CES Common Consent dataset. Even though it had over 61,000 respondents, over 6,000 respondents either didn't register to vote or had no idea if they were registered at all. Moreover, after data cleaning was completed, we were left with 43,540 entries in the dataset - over a third of the data was lost. Furthermore, in these 43,540 entries, the ratio of male to female entries weighed greatly in favor of females, and was much higher than the national gender ratio in 2020.

Due to the nature of the survey being online, a significant portion of the US's demographics was left out of the pool. It is estimated that in 2020 about 13% of the US population did not have access to the internet (Petrosyan (2024)), which leads to the exclusion of about 43 million citizens being excluded from the survey with no representation at all.

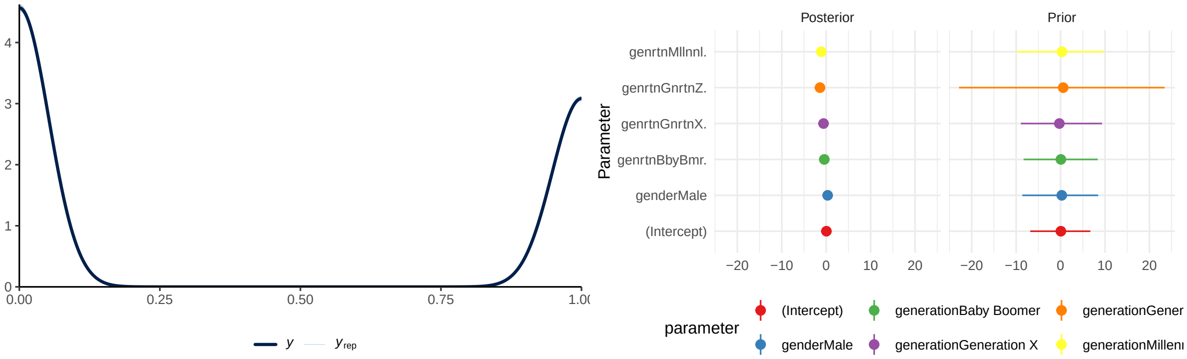
It would be interesting to study how different generations tend to vote in the previous US presidential elections (which occurred under more normal circumstances) as well, especially the cases where a new generation was added to the voter pool of the nation, similar to 2020. Another interesting avenue would be to study voting trends amongst women over the last few decades, as the scope of this paper limits us from exploring exactly how great of a shift occurred in 2017.

Appendix

A Model details

A.1 Posterior predictive check

In Figure 5a we implement a posterior predictive check. In Figure 5b we compare the posterior with the prior.



(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 5: Examining how the model fits, and is affected by, the data

A.2 Diagnostics

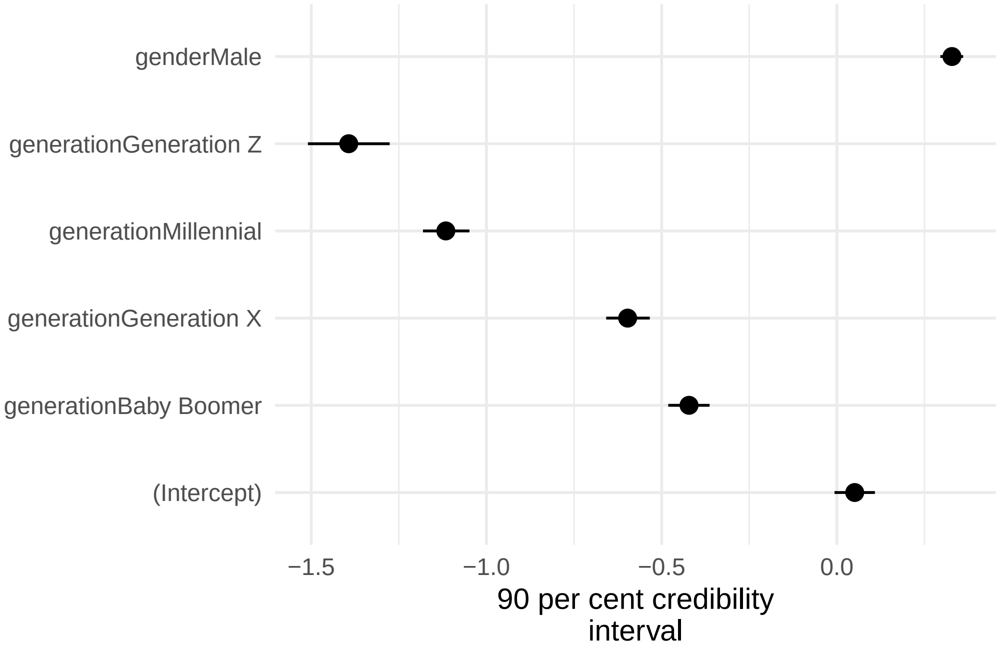
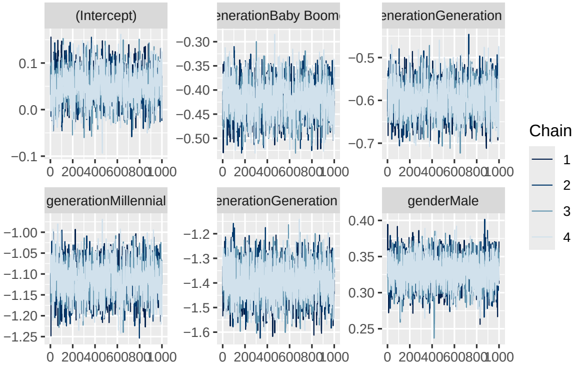


Figure 6: Credible Intervals



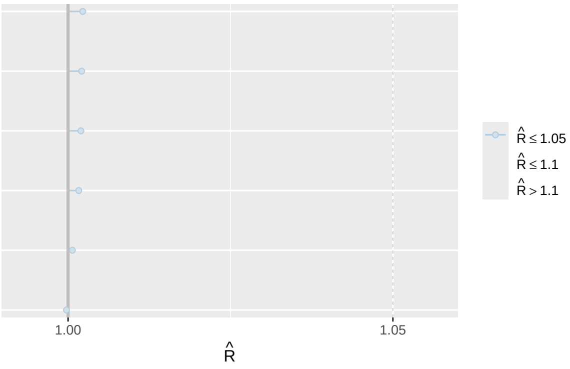


Figure 7: Trace plot

Figure 8: Rhat plot

References

2020 Presidential Election Voting and Registration Tables Now Available. 2021. United States Census Bureau. https://www.census.gov/newsroom/press-releases/2021/2020-presidential-election-voting-and-registration-tables-now-available.html.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Dimock, Michael. 2019. "Defining Generations: Where Millennials End and Generation z Begins." https://www.pewresearch.org/short-reads/2019/01/17/where-millennials-end-and-generation-z-begins/.

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

———. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Gramlich, John. 2020. "What the 2020 Electorate Looks Like by Party, Race and Ethnicity, Age, Education and Religion." https://www.pewresearch.org/short-reads/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion/.

Hess, Abigail Johnson. 2020. "The 2020 Election Shows Gen z's Voting Power for Years to Come." https://www.cnbc.com/2020/11/18/the-2020-election-shows-gen-zs-voting-power-for-years-to-come.html.

Itkowitz, Colby. 2019. "The Next Generation of Voters Is More Liberal, More Inclusive and Believes in Government." https://www.washingtonpost.com/politics/2019/01/17/next-generation-voters-are-more-liberal-more-inclusive-believe-government/.

Jurkowitz, Mark. 2020. "Majority of Americans Disapprove of Trump's COVID-19 Messaging, Though Large Partisan Gaps Persist." https://www.pewresearch.org/short-reads/2020/09/15/majority-of-americans-disapprove-of-trumps-covid-19-messaging-though-large-partisan-gaps-persist/#:~:text=The%20survey%20also%20finds%20large, say%20it%20is%20completely%20wrong.

Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *Dataverse: R Client for Dataverse 4+ Repositories.*

Petrosyan, Ani. 2024. "United States Internet Penetration 2000-2024." https://www.statista.com/statistics/209117/us-internet-penetration/.

R Core Team. 2023. *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://CRAN.R-project.org/package=arrow.

Rosie Campbell, Rosalind Shorrocks. 2023. "Women Used to Be More Likely to Vote Conservative Than Men but That All Changed in 2017 – We Wanted to Find Out Why." https://theconversation.com/women-used-to-be-more-likely-to-vote-conservative-than-men-but-that-all-changed-in-2017-we-wanted-to-find-out-why-214019.

Saad, Lydia. 2024. "U.s. Women Have Become More Liberal; Men Mostly Stable." https://news.gallup.com/poll/609914/women-become-liberal-men-mostly-stable.aspx.

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

———. 2021a. "Cooperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.

"The Gender Ratio of United States of America (2020 - 2028, Males Per 100 Females)." 2024. GlobalData. https://www.globaldata.com/data-insights/macroeconomic/the-gender-ratio-of-united-states-of-america-325511/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.