Income and Gender: Forecasting the U.S. Voters' Choices in 2018*

A Statistical Analysis Explores the Correlation Between Socioeconomic Indicators and Electoral Decisions

Siqi Fei

March 9, 2024

In this study, we examined whether it is possible to predict an individual's voting behaviour in the 2018 U.S. election based solely on their income level and gender. By utilizing a dataset containing voter information and demographic characteristics, we conducted statistical analyses to identify any patterns or correlations.Our findings suggest that, to a certain extent, a person's income level and gender can provide predictive insights into their voting preferences, shedding light on the complex interplay between socio-economic factors and political affiliation. Understanding these relationships can help policymakers and political strategists craft more targeted approaches to engage voters across different demographic groups, ultimately contributing to a more informed and refined electoral system.

Table of contents

1 Introduction **2**

2 Data **3**

2.1 Data Measurement 3

2.2 Data Distribution 4

3 Model **5**

3.1 Model set-up 5

3.1.1 Model Justification 6

4 Results **7**

*Code and data are available at: https://github.com/FXXFERMI/Political-support-in-the-United-States.git.

5 Discussion **9**
5.1 Overview 9
5.2 Gender Influence on Voting Patterns 9
5.3 The Role of Annual Family Income 9
5.4 Limitations 10
5.5 Future Research 10
5.6 Conclusion 11

Appendix **12**

A Model details **12**

References **13**

1 Introduction

As the largest political center in the world, U.S. presidential election is heavily influenced by various sectors including social, cultural, political, and economic factors. This article conducts an in-depth investigation on the 2018 U.S. election, focusing on the impact of annual family income level and gender on electoral choices. The 2018 election was chosen based on the high voter participation rate and clear partisan divisions. According to a Pew Research Center analysis (Pew Research Center 2023), 2018 election has 49% turnout which is the highest rate for a midterm since 1914. Adults who voted in at least one election during the period divide evenly between Democrats and independents who lean toward the Democratic Party or Republicans and Republican-leaning independents in their current party affiliation (48% each). The active engagement and high level of discussion also allowed our dataset to cover a wider range of individuals and societal characteristics.

This study uses data from the Cooperative Congressional Election Study (CCES) (Schaffner, Ansolabehere, and Luks 2019), employing Logistic Regression and quantitative methods to explore the relationship between voters' income level and gender and their support for political party candidates. The primary estimand of our analysis is the likelihood of voting for a specific political party or candidate based on an individual's income level and gender. Our research shows that these two factors significantly influence voter preference, providing a deeper understanding of the political inclinations of different demographic structures. This study enriches the academic theory of election prediction and has practical significance for developing more detailed and effective election strategies. By identifying gender and income as important determinants of voting patterns, we lay the groundwork for future investigations into how these and other demographic factors shape political identity and loyalty. Furthermore, this research helps us gain a more comprehensive understanding of the political landscape in the United States and enhances our understanding of global politics.

The paper is organized into four sections, following this introduction. Section 2 describes the dataset and the rationale behind its selection. Section 3 outlines the statistical methods applied to analyze the data. Section 4 presents the key findings of the study, while Section 5 interprets these results within the broader context of American politics and voting behaviour. Through this structure, we aim to offer a comprehensive examination of how income and gender influence voting decisions in the United States, contributing to the broader discourse on democracy and electoral participation.

2 Data

For this paper, we used data from the 2018 Cooperative Congressional Election Study (CCES) (Schaffner, Ansolabehere, and Luks 2019), which we worked with in R (R Core Team 2023), a language for statistical computing. The `tidyverse` suite (Wickham et al. 2019), with its various packages like `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), and `tibble` (Müller and Wickham 2023), made handling the data easier and more precise. We summarized our model results using the `modelsummary` package (Arel-Bundock 2022). The `dataverse` package (Kuriwaki, Beasley, and Leeper 2023) helped us smoothly download the data, and `testthat` (Wickham 2011) made sure our data prep and analysis were solid. The `here` package (Müller 2020) helped keep our files organized and our analysis reproducible.

2.1 Data Measurement

Our study focuses on three main data points: annual family income, gender, and individual voting choices in the 2018 election. We organize income into 12 ranges to examine the potential effect of earnings on political decisions. Gender is categorized simply into male or female, reflecting the dataset's binary format. Regarding voting choices, we concentrate on whether respondents voted for Trump or Clinton, framing it as a binary outcome.

There are 29,756 observations in the raw dataset, before analyzing, we thoroughly clean our data by removing non-essential information and align income brackets with recognized standards. 18,279 observations remains in our study. We also took deliberate steps to address data quality:

- We excluded all respondents with 'NA' in the income data field. This decision was made to ensure the integrity of our socioeconomic analysis and to maintain consistency in the income-related variables, which are central to our study.
- We limited our analysis to participants who voted for either Trump or Clinton, excluding those who voted for other candidates. This was done to focus on the two primary candidates, thereby simplifying the analysis and ensuring a clear binary choice reflective of the major party candidates.

- It is important to note that non-binary individuals were not included in the gender analysis of this study. This was due to the dataset's limitations, which did not account for non-binary gender identities. While this is a recognized gap, addressing it would require a different dataset and potentially a distinct analytical approach, which could be considered for future research to provide a more inclusive understanding of voting behavior.

By implementing these measures, we aimed to create a robust and reliable dataset that allows for a focused examination of the influence of income level and binary gender on voting behavior in the pivotal 2018 election.

2.2 Data Distribution

We present a figure to illustrate our data's narrative. Figure 1 depicts the voting choices broken down by gender and annual family income levels. This bar chart provides a clear visualization of the number of respondents within each income category and their voting behaviour.

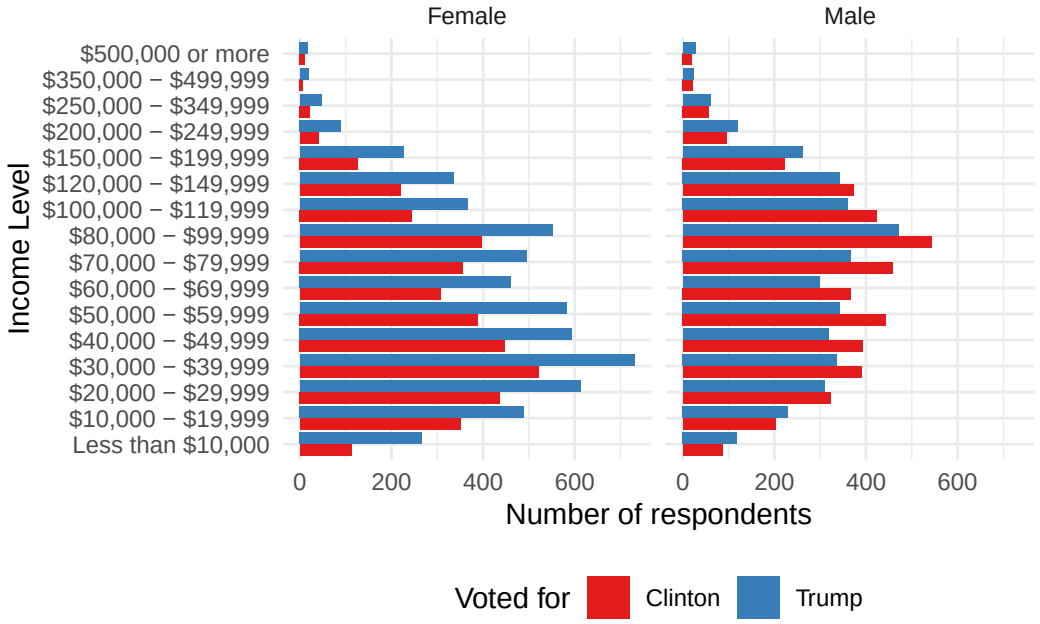

Figure 1: The distribution of presidential preferences, by gender, and annual family income

The bar chart is divided into two major sections: one for female respondents and the other for male respondents. Within each gender category, there are multiple income ranges represented on the vertical axis. The horizontal axis represents the number of respondents within each income bracket who voted for a particular candidate- red for Clinton and blue for Trump.

From reviewing the summary statistics, we notice distinct patterns. For example, higher-income males predominantly supported Trump, females across most income brackets showed significant support for Clinton. This could indicate gender-aligned priorities or responses to the candidates' platforms. Also, it appears that income level has a varied influence on voting patterns for Clinton but a more consistent pattern for Trump, particularly among males. Lower-income brackets do not show as strong of a preference for Trump as higher-income brackets do. Moreover, middle-income voters, particularly among males, show a balance in their support for the two candidates. This could reflect a group of voters who may be more moderate or whose voting decisions are influenced by a mix of factors beyond just income.

We also noticed an interesting divergence in high-income females, where support for Clinton is not as strong as it is among women in some lower-income brackets. This might suggest that at higher-income levels, the gender based voting pattern is less pronounced.

In simpler terms, the chart suggests that a voter's income and gender might have played a role in who they choose to vote for. Higher earners, especially men, leaned towards Trump, while women across various income levels showed considerable support for Clinton. The two middle-income groups were more mixed, showing no strong preference for neither candidate, which might indicate that factors other than income influenced their choices.

3 Model

The goal of our modeling strategy is to illuminate the underlying dynamics between voters' demographics—specifically, income and gender—and their voting behavior in the 2018 U.S. elections. Additionally, we aim to quantify the strength and direction of these relationships.

To dissect these factors, we apply a Bayesian analysis model, allowing us to incorporate prior knowledge and handle uncertainty in a probabilistic framework. We chose this approach due to its flexibility and the rich interpretability of its results. For those interested in the technical specifics and diagnostic checks of our Bayesian model, these details are thoroughly presented in Appendix A.

3.1 Model set-up

We denote y_i as the binary outcome where 1 represents a vote for Trump and 0 represents a vote for Clinton. In our model, the predictors include gender and income. We use β_1 to represent the effect of gender on the probability of voting for Trump, with gender encoded as a binary variable (e.g., 0 for female and 1 for male). Similarly, β_2 represents the effect of income, categorized into distinct income brackets, on the voting probability.

$$y_i|\pi_i \sim \text{Bernoulli}(\pi_i) \tag{1}$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{income}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

The parameters β_0,β_1,β_2 represent the intercept and the effects of income and gender, respectively, on the log-odds of voting for Trump. A normal prior with mean 0 and standard deviation 2.5 is used for each of these parameters, reflecting a baseline assumption of no effect before observing the data.

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

Given the economic and social context of the 2018 election, it is reasonable to hypothesize that both income and gender would exert influence on voting decisions. We anticipate that income, as a measure of economic status, could have a significant effect on the choice of a candidate known for advocating policies affecting taxations and business. Similarly, gender might play a role given the differing policy stances of the candidates on issues traditionally seen as gendered in public discourse. Therefore, we posit a positive relationship between higher income brackets and support for Trump, while also considering the potential impact of gender on voting behavior.

The subsequent results section will delve into the outcomes of this Bayesian model, revealing the extent to which income and gender can predict support for Trump in the 2018 U.S. elections.

4 Results

The results of our analysis, grounded in the 2018 Cooperative Congressional Election Study (CCES)(Schaffner, Ansolabehere, and Luks 2019), offer a compelling narrative about the interplay between income, gender, and voter preference in the 2018 U.S. elections.

Our results, as summarized in Table 1, explain the relationship between an individual's likelihood to support Trump in the 2018 election and their gender and annual family income level. The intercept, representing the baseline likelihood for the reference group (here presumably female and the omitted income category), is estimated at 0.44. Gender presents a significant effect, with males being less likely by an odds ratio of 0.50 to support Trump, which corresponds to a 50% decrease in the odds, holding other factors constant.

The impact of annual family income on voting preference for Trump is intricate. For most of the income brackets, the coefficients are not substantial in magnitude and lack statistical significance, considering their respective standard errors. Yet, individuals earning between \$200,000 to \$249,999 show a noticeably higher likelihood (an increase of 0.31) of supporting Trump. Similarly, the highest income bracket of \$500,000 or more also suggests a higher likelihood, with a coefficient of 0.28. In a compelling contrast, individuals in the lowest income bracket, earning less than \$10,000, demonstrate a likelihood increase of 0.40 to support Trump, opposing some conventional assumptions about income and political preference.

The \$R^2\$ value of 0.019 implies that, though gender and income are predictive of voting behavior, a significant amount of variance remains unexplained by these variables alone. It is probable that other, unobserved variables also significantly influence an individual's voting choices.

The model's fit, assessed by criteria such as the Widely Applicable Information Criterion (WAIC) and the Leave-One-Out Cross-Validation Information Criterion (LOOIC), indicates that the model is aptly calibrated given the complexity of human voting behavior it seeks to capture. Nonetheless, the modest R-squared value suggests that incorporating additional predictors might enrich future research, offering a more intricate understanding of the factors that drive voter behavior.

In summary, the analysis confirms that gender and annual family income are indeed predictive of voter preference, yet the variability indicates that additional factors also play a significant role.

Table 1: Whether a respondent is likely to vote for Trump based on their gender and annual family income

	Support Trump
(Intercept)	0.44
	(0.05)
genderMale	−0.50
	(0.03)
income$100,000 - $119,999	−0.07
	(0.08)
income$120,000 - $149,999	−0.02
	(0.08)
income$150,000 - $199,999	0.19
	(0.09)
income$20,000 - $29,999	−0.05
	(0.08)
income$200,000 - $249,999	0.31
	(0.12)
income$250,000 - $349,999	0.20
	(0.15)
income$30,000 - $39,999	−0.09
	(0.07)
income$350,000 - $499,999	0.28
	(0.25)
income$40,000 - $49,999	−0.15
	(0.07)
income$50,000 - $59,999	−0.10
	(0.07)
income$500,000 or more	0.28
	(0.23)
income$60,000 - $69,999	−0.08
	(0.07)
income$70,000 - $79,999	−0.13
	(0.07)
income$80,000 - $99,999	−0.09
	(0.07)
incomeLess than $10,000	0.40
	(0.11)
Num.Obs.	18 279
R2	0.019
Log.Lik.	−12 442.206
ELPD	−12 459.1
ELPD s.e.	21.2
LOOIC	24 918.1
LOOIC s.e.	42.4
WAIC	24 918.1
RMSE	0.49

5 Discussion

5.1 Overview

In this study, we explored the possibilities that annual family income and gender might influence voting behaviour, focusing on data from the 2018 U.S. election. We used Logistic Regression and Bayesian analysis to delve into the probabilities and patterns that suggest whether certain demographics groups were more likely to vote for Trump. This approach gives us a deeper statistical look into the elements that sway electoral decisions.

5.2 Gender Influence on Voting Patterns

The study has shed light on gender as a predictor of voting preferences. Statistically speaking, males showed a lower probability of voting for Trump, which echoes the gendered divide widely reported in political analyses. Interestingly, this finding challenges some traditional assumptions about male voting patterns. The long-held view is that men tend to favor Republican candidates due to factors like economic conservatism (Smith 2016). Our finding suggests that gender may have a more nuanced impact on political preferences than previously assumed. It hints that men's voting preference in the 2018 election could have been influenced by a host of factors, possibly including the candidates' policies on economic and social issues that resonate differently with men and women.

Moreover, the gender based voting patterns we identified may reflect broader cultural shifts or reactions to the political climate at the time. The numbers point to an evolving political identity among both men and women, revealing that gender is not just a demographic detail but a lens through which political platforms and candidate personas are evaluated.

5.3 The Role of Annual Family Income

When we turn to income level, the story becomes more intricate. Our statistical approach showed that support for Trump spanned across the income spectrum, challenging the typical narrative that assumes a straight-line correlation between income level and conservative voting. Notably, the trend does not hold to a simple high-income-conservative, low-income-liberal dichotomy. Instead, it reveals that individuals from both ends of the family income scale found appeal in Trump's candidacy.

This complex relationship might suggest that for some voters, personal financial considerations intersect with broader economic ideologies in unexpected ways. For example, high-income voters might support a candidate they believe will protect their financial interests, while low-income voters might be swayed by promises of economic change or protectionist policies. Alternatively, it could indicate that for many voters, non-economic issues such as

cultural identity, party loyalty, or candidate character play a more pivotal role in their voting preference than their income bracket would suggest.

5.4 Limitations

The scope of our study is not without limitations. Primarily, the explanatory power of our model indicated by a low R-squared value, suggests that the variables of annual income and gender do not capture the full spectrum of factors influencing voters' decisions.

A notable limitation is the binary classification of gender used in our analysis. This classification does not account for non-binary and transgender individuals, whose political preferences and voting behaviors may differ significantly from those within the binary gender framework. Additionally, the broad income brackets used in the study may obscure the subtleties of socioeconomic status and its diverse impacts on voting patterns.

Our cross-sectional data provides a valuable snapshot but falls short of explaining longitudinal trends or establishing causality. The dynamic nature of voter behavior, influenced by rapidly evolving social and economic contexts, calls for more nuanced data collection methods.

5.5 Future Research

Future research should aim to include a more comprehensive array of variables, such as ethnic background, educational attainment, and geographic factors, to provide a richer statistical portrait of voter behavior. Longitudinal research would be invaluable in tracing how these patterns shift over time and in understanding the deeper currents that shape electoral outcomes.

In-depth qualitative methods, such as interviews or focus groups, could be employed to supplement our quantitative analysis. Such approaches would allow for a more profound exploration of individual motivations and the complex emotional and cognitive processes that guide voters at the polls.

Moreover, subsequent studies should make concerted efforts to embrace gender diversity by moving beyond binary gender classifications to reflect a more nuanced and inclusive understanding of gender identity. In terms of economic indicators, it would be beneficial to analyze finer gradations of income and to consider other markers of economic status, such as wealth, debt, and financial stability, which may offer additional insights into the interplay between economic standing and voting behavior.

5.6 Conclusion

This paper is a contribution to the broader puzzle of what influences voters in their electorial decisions. By unraveling the statistical significance of demographics like annual family income and gender, we open up avenues for richer political dialogue and more strategic campaign planning. Each study, including ours, adds another layer to our collective understanding of the American electorate, inching us closer to demystifying the complexities of voter behaviour.

Appendix

A Model details

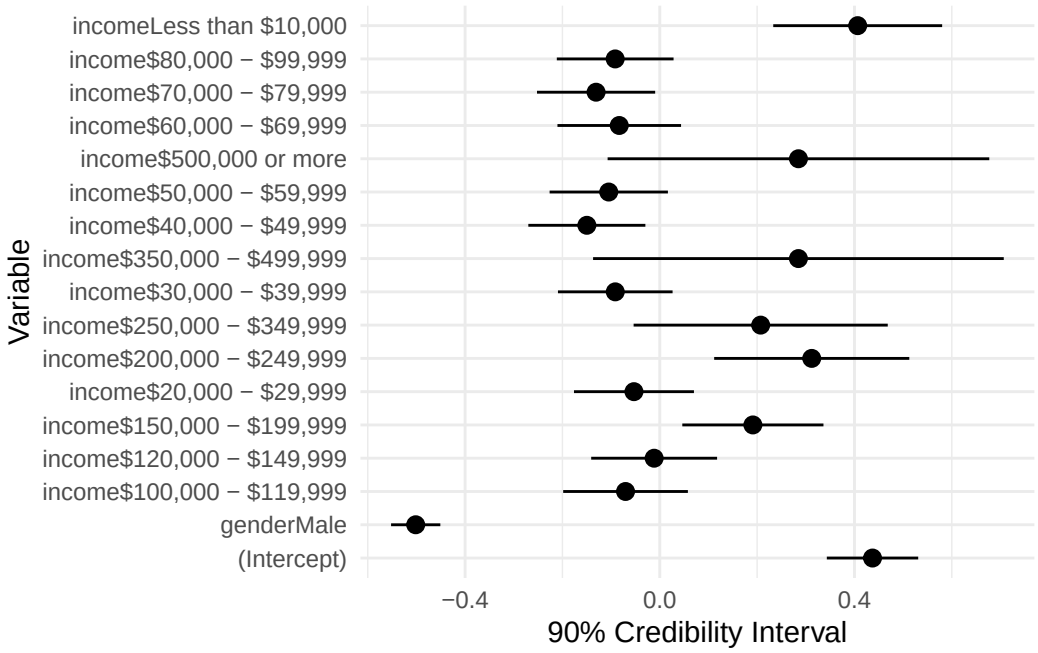

Figure 2: Credible intervals for predictors of support for Trump

References

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.

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

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

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Müller, Kirill, and Hadley Wickham. 2023. *tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Pew Research Center. 2023. "Voter Turnout 2018-2022." July 12, 2023. https://www.pewresearch.org/politics/2023/07/12/voter-turnout-2018-2022/.

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

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2019. "CCES Common Content, 2018." Harvard Dataverse. https://doi.org/10.7910/DVN/ZSBZ7K.

Smith, A. 2016. *Gender and Voting Behavior in American Elections*. New York: Horizon Press.

Wickham, Hadley. 2011. "testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

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

Wickham, Hadley et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

——— et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.