# Age, Education, Income, and the Canadian Vote. Understanding the 2021 election.*

Allen Uy

March 12, 2024

This study examines the factors shaping support for Canada's major parties, the Liberal and Conservative, during the 2021 federal election. Analyzing personal and topical variables, we discovered that age has minimal impact, higher education correlates with increased Liberal support, and higher income is linked to Conservative favor. These insights offer a comprehensive understanding of the 2021 election dynamics, enabling more informed predictions for the upcoming 2025 election. This research contributes to a better grasp of Canadian political affiliations, assisting citizens and policymakers in adapting to evolving electoral trends.

## Table of contents



*Code and data are available at: https://github.com/varygx/CES2021.

# 1 Introduction

The two largest political parties in Canada, the Liberal Party and Conservative party won 160 and 119 out of 338 seats respectively in the 44th federal election held in 2021 (Canada (2021)). With the 45th federal election approaching in 2025, this paper conducts an analysis on the political support for these two parties, specifically examining personal and topical variables. The estimand of interest in this research is the average causal effect of these variables on the likelihood of individuals expressing support for either the Liberal or Conservative party.

By exploring personal and topical variables, this paper aims to provide insight into the factors that influenced political support for the Liberal and Conservative parties in Canada. The goal is to provide a clearer understanding of the dynamics that shaped the 2021 election and offer insights into broader patterns of political affiliations in the Canadian context, which could be used for inference in the upcoming election.

The subsequent sections follow a structured format. The Data section outlines the source and variables central to our analysis. The Model section details the construction and methodology of the statistical models used. The Results section presents the key findings of our analysis, while the Discussion section critically reviews the content, addresses the implications of the results, acknowledges model limitations, and suggests potential future research directions.

# 2 Data

The data used in this paper was gathered from the 2021 Canadian Election Study (CES) hosted on the Harvard Dataverse (Stephenson et al. (2022)) and analyzed using R (R Core Team (2023)) with help from `tidyverse` (Wickham et al. (2019)), `haven` (Wickham, Miller, and Smith (2023)), `arrow` (Richardson et al. (2024)), `rstanarm` (Goodrich et al. (2022)),

Table 1: Sample of Personal Variable Data

| age | education | income_cat |
|---|---|---|
| 57 | Some technical, community college, CEGEP, College Classique | $30,001 to $60,000 |
| 29 | Master's degree | $60,001 to $90,000 |
| 66 | Completed secondary/ high school | $110,001 to $150,000 |
| 42 | Completed secondary/ high school | $60,001 to $90,000 |
| 33 | Completed technical, community college, CEGEP, College Classique | $110,001 to $150,000 |

`modelsummary` (Arel-Bundock (2022)), `testthat` (Wickham (2011)), `here` (Müller (2020)), `knitr` (Xie (2023)), and `kableExtra` (Zhu (2021)).

## 2.1 CES2021

The dataset was gathered by CES through an online survey via the Leger Opinion platform where data was finalized on November 1, 2023. A total of 20968 responses were recorded with the stratification goals contained in Appendix A.

The survey was launched during what they defined as the Campaign Period Survey (CPS) during August 17 to September 19, 2021 and a follow-up Post-Election Survey (PES) had 15069 responses during September 23 to October 4, 2021.

The provided data was cleaned by CES to include mostly high-quality responses, the details of which can be found in Appendix B.

The dataset includes 1059 variables, many of which could have been included in the analysis but was narrowed down to 30 variables that could correlate with party support.

These variables are gathered from the CPS portion of the survey with no open-ended answers and assigned numerical values with labels.

### 2.1.1 Personal Variables

Age was calculated in years based on the year of birth the respondent inputted. Education, Employment, Religion, Immigration Status, and Province were given a number with a corresponding label based on which radio button the respondent selected. Income category was given a number corresponding to a range of household incomes. Respondents were asked to input their income or if they felt uncomfortable to only specify the range. Exact income numbers were converted to their appropriate range. A preview of personal variables can be seen in Table 1.

Table 2: Sample of Topic Variable Data

| econ_party | healthcare_party | imm_party | covid_party |
|---|---|---|---|
| Conservative Party | Conservative Party | ndp | Conservative Party |
| NA | NA | NA | NA |
| Conservative Party | Conservative Party | Conservative Party | Conservative Party |
| Conservative Party | Conservative Party | Conservative Party | Conservative Party |
| Liberal Party | Liberal Party | Liberal Party | Liberal Party |

Table 3: Sample of the Party the Respondent Supports

| voted_for |
|---|
| Conservative |
| Liberal |
| Conservative |
| Conservative |
| Liberal |

### 2.1.2 Topic Variables

The following survey topics were extracted from the dataset: Economy, Healthcare, Immigration, COVID, Environment, Quebec Sovereignty, Government Spending, Education, Housing, Carbon Tax, and Childcare. Variables ending in `_party` indicate the party which the respondent believes would handle the best. The survey was designed to give only half the surveyees this question. Other variables are from questions with responses such as: "(strongly) disagree, neutral, (strongly) agree" where the respondent selected one of those answers or did not answer. A preview of topic variables can be seen in Table 2.

### 2.1.3 Party Variables

Depending on the surveyee's circumstances, the survey asked which party they voted for in advance, will vote for, are likely to vote for, or would vote for if possible. Only one of these questions was ever asked so we aggregated those columns into a single `voted_for` variable, filtered to the Liberal and Conservative parties. A preview of the `voted_for` variable can be seen in Table 1.

## 3 Model

We investigate model that might explain political support. A logistic regression model using age, education, and income.

### 3.1 Model set-up

Define $y_i$ as the political preference of the respondent and equal to 1 if Liberal and 0 if Conservative. Then age, education, and income are the respective answers of the respondent.

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \times \text{age}_i + \beta_2 \times \text{education}_i + \beta_3 \times \text{income}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

We assume Normal priors.

## 4 Results

Based on Table 4 we observe that age has little effect on party support. It also suggests those with higher education are likely to vote Liberal while those with higher income are more likely to vote Conservative.

## 5 Discussion

### 5.1 Lack of Topical Model

The paper omits a model based on topic variables due to compute feasibility. Due how the data was stored, each variable is a number that was converted to a factor using `as_factor` from `haven` (Wickham, Miller, and Smith (2023)). These variables would often have 3 to 5 levels for each factor or more. This lead to large models for the topic model and even the personal model. This was simply not feasible, however the data processing prior to the decision to omit the topic model and simplify the personal model was kept in the paper.

### 5.2 Model Limitations

### 5.3 Next Steps

Table 4: Explanatory model of party support based on age, education, and income

| | Personal model |
|---|---|
| (Intercept) | −1.88 |
| | (1.60) |
| age | −0.01 |
| | (0.00) |
| educationSome elementary school | 2.74 |
| | (1.79) |
| educationCompleted elementary school | 2.05 |
| | (1.75) |
| educationSome secondary/ high school | 2.15 |
| | (1.65) |
| educationCompleted secondary/ high school | 2.40 |
| | (1.64) |
| educationSome technical, community college, CEGEP, College Classique | 2.41 |
| | (1.61) |
| educationCompleted technical, community college, CEGEP, College Classique | 2.43 |
| | (1.63) |
| educationSome university | 2.75 |
| | (1.64) |
| educationBachelor's degree | 2.95 |
| | (1.63) |
| educationMaster's degree | 3.26 |
| | (1.62) |
| educationProfessional degree or doctorate | 3.06 |
| | (1.63) |
| educationDon't know/ Prefer not to answer | 2.16 |
| | (1.84) |
| income_cat$110,001 to $150,000 | −0.45 |
| | (0.08) |
| income_cat$150,001 to $200,000 | −0.65 |
| | (0.10) |
| income_cat$30,001 to $60,000 | −0.19 |
| | (0.08) |
| income_cat$60,001 to $90,000 | −0.33 |
| | (0.07) |
| income_cat$90,001 to $110,000 | −0.54 |
| | (0.09) |
| income_catDon't know/ Prefer not to answer | −0.29 |
| | (0.15) |
| income_catMore than $200,000 | −0.65 |
| | (0.11) |
| income_catNo income | −0.61 |
| | (0.14) |
| Num.Obs. | 10 525 |
| R2 | 0.031 |
| Log.Lik. | −7124.288 |
| ELPD | −7145.2 |
| ELPD s.e. | 18.2 |
| LOOIC | 14 290.4 |
| LOOIC s.e. | 36.5 |
| WAIC | 14 290.3 |
| RMSE | 0.49 |

# Appendix

## A Stratification Goals

### A.1 Gender

- 50% men
- 50% women

### A.2 Age

- 28% aged 18-34
- 33% aged 35-54
- 39% aged 55 and higher

### A.3 Region

- 7% in the Atlantic
- 23% in Quebec
- 38% in Ontario
- 32% in the West

### A.4 Language

- 80% French in Quebec, 20% English
- 10% French within the Atlantic
- 10% French nationally

## B Response Quality

As per Stephenson et al. (2022), responses were kept according to the following on page 6 of the codebook:

During the data cleaning process, respondents were categorized based on their most important reason for removal. While respondents might be removed for multiple reasons, the most important reason is most relevant. Reasons for removal are, in order of importance:

1. Internal survey testing or previews
2. Ineligible - did not consent to survey
3. Ineligible - not a Canadian citizen or permanent resident
4. Ineligible - respondent under 18 years of age
5. Over quota
6. Incomplete - did not complete the survey
7. Duplicate of previous respondent (identified by survey panel ID)
8. Duplicate of a previous respondent (identified by IP address and the following demographics: year of birth, gender, education level, employment, religion, immigration status)
9. Speeder (completed the survey in less than 500 seconds, or 8.3 minutes)
10. Postal code-province mismatch
11. Straightliner
12. Failed attention check
13. YOB mismatch
14. Province mismatch
15. Inattentive
16. Initial duplicate (identified by survey panel ID)

The following values are in the final dataset

17. Initial duplicate (identified by IP address and demographics)
18. PES speeders (respondents that took between 6 and 7.12 minutes to complete the PES)
19. Clean complete

## 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.

Canada, Elections. 2021. "FORTY-FOURTH GENERAL ELECTION 2021 Official Voting Results." https://elections.ca/res/rep/off/ovr2021app/53/table7E.html.

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

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

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.

Stephenson, Laura B, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2022. "2021 Canadian Election Study (CES)." Harvard Dataverse. https://doi.org/10.7910/DVN/XBZHKC.

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.

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, Evan Miller, and Danny Smith. 2023. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://CRAN.R-project.org/package=haven.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.