

Analysing social factors influence Canadian voters' choices*

Decoding CES 2021 Vote Data by using A Multilevel Logistic Regression Model

Yichen ji

April 2, 2024

1) what was done, 2) what was found, and 3) why this matters (all at a high level). .

Table of contents

1	Introduction	2
1.1	estimand	3
2	Data	3
2.1	Analysis Tools	3
2.2	Variables	3
2.3	Sample of cleaned data	5
2.4	Survey data summary	5
2.5	Measurement	5
3	Model	7
3.1	Multilevel Logistic Regression Model	7
3.1.1	Model justification	7
4	Results	9
4.1	Model Coefficients	9
5	Discussion	9
5.1	What is done in this paper?	9
5.2	What is something that we learn about the world?	9

*Code and data are available at: https://github.com/Selinayichenji/Canadian_election_prediction.git.

5.3	What is another thing that we learn about the world?	9
5.4	Weaknesses and next steps	9
A	Appendix	11
B	Additional data details	11
C	Model details	11
C.1	Posterior predictive check	11
C.2	Diagnostics	11
	References	13

1 Introduction

This report utilizes the 2021 Canadian Election Study data and the “General Social Survey, Cycle 31: Families, Public Use Microdata File Documentation and User’s Guide” (Statistics Canada, 2019) to solve the emergent dilemma of diminishing vote participation rate. The primary purpose of this report is to use the data set to predict future voter turnout by differentiating throughout the province(note: not include Quebec, Yukon, Northwest Territories, and Nunavut).

Based on previous voting analysis, the main factors affecting voting include age, education level, family status(represented by the number of children), and economic well-being(represented by the family income). (“Factors Associated with Voting,” 2015) Thus, the examination will focus on demographic variables, including age, gender, family income, education level, and the number of children in the vote rate corresponding to Canada’s three main political parties. The three main political parties are Liberal, Conservative, and NDP(Canada’s New Democrats), which occupied around 90 percent of the election 2010s. (“Canadian Federal Election Results,” n.d.) The report will use logistic regression on different parties to model the participation rate. Whether the voting for particular parties or not will be recorded binary, as 1 or 0. Under this method, we assume the authenticity of our source, 2021 Canadian Election Study data, to build up the model.

The importance of this study is to reveal the complexities behind the recent decline in voter participation rates. By exploring the complex demographic factors and the voting rate of different political parties, this report could provide insights into the advantages of each of the three main parties during the election. According to past data, where liberals always take up the majority of the voting, we will assume the liberals will win the overall election. (Simon Fraser University, n.d.) Meanwhile, this report will examine how various demographic elements correlate with political affiliations, aiming to understand the challenges facing Canadian democracy comprehensively. In addition, the findings can not only predict the democratic participation rate throughout different states but also serve as a potential solution to the diminishing voting

participation rate corresponding to each province, which can strengthen democracy throughout the nation.

1.1 estimand

The paper introduces the basic information contains data source, methodology, variables and measurements in Section 2. The visualization and analysis of the tendency differences in moderation by people across three social factors are presented in Section 4. And in Section 5, we discuss what we have learned, our understanding of the world, the limitations, and the next steps of our research. Section A is for appendix and Section C.2 is listed all references in this paper.

2 Data

The survey data was sourced from the “2021 Canadian Election Study” (Stephenson, Harell, Rubenson, & Loewen, 2022).

2.1 Analysis Tools

Our paper applies using the statistical programming language R (R Core Team 2023). Besides the programming tool, we also employ the following packages: readr (?), ggplot2 (?), dplyr (?), tidyverse (?), MetBrewer (?), knitr (?), tidyr(?), kableExtra(?) and grid(?).

2.2 Variables

Age: include five different groups: 18-29 years old, 30-44 years old, 45-59 years old, 60-74 years old, and above 74 years old. We manipulate the age for census data, which has different years than the survey data.

vote_liberal/vote_conservative/vote_NDP: the voting result on whether to vote for three different parties. (Liberal, Conservative, or NDP) The result of voting or not is recorded in binary.

Gender: only include two types of gender, female and male, excluding non-binary or others

education_level: including six different education levels(Less than high school, High school, Non-University, University certificate below the bachelor, Bachelor’s degree, Above the bachelor).

family_income: including six different groups. (“Less than 25,000”, “25,000 to 49,999”, “50,000 to 74,999”, “75,000 to 99,999”, “100,000 to 124,999”, “125,000 and more”)

Province: 9 different states and excluding Quebec, Yukon, Northwest Territories, and Nunavut.

children_number: the number of children within the family, which includes only six levels (0,1,2,3,4 and 4+ children)

Table 1: Sample of the Cleaned Data

age	vote_liberal	vote_conservative	vote_NDP	gender	education_level	family_income	province	children_number
18-29	0	0	1	Female	University certificate below the bachelor	\$125,000 and more	British Columbia	0
45-59	0	0	1	Female	Above the bachelor	\$125,000 and more	Ontario	2
45-59	0	0	1	Female	University certificate below the bachelor	\$50,000 to \$74,999	Ontario	1
60-74	0	1	0	Male	University certificate below the bachelor	\$100,000 to \$124,999	British Columbia	1
45-59	1	0	0	Male	Non-University	\$50,000 to \$74,999	Manitoba	0

2.3 Sample of cleaned data

2.4 Survey data summary

All detailed tables include summary data of 6 variables are in Section A age, gender, income, education level, province, children number

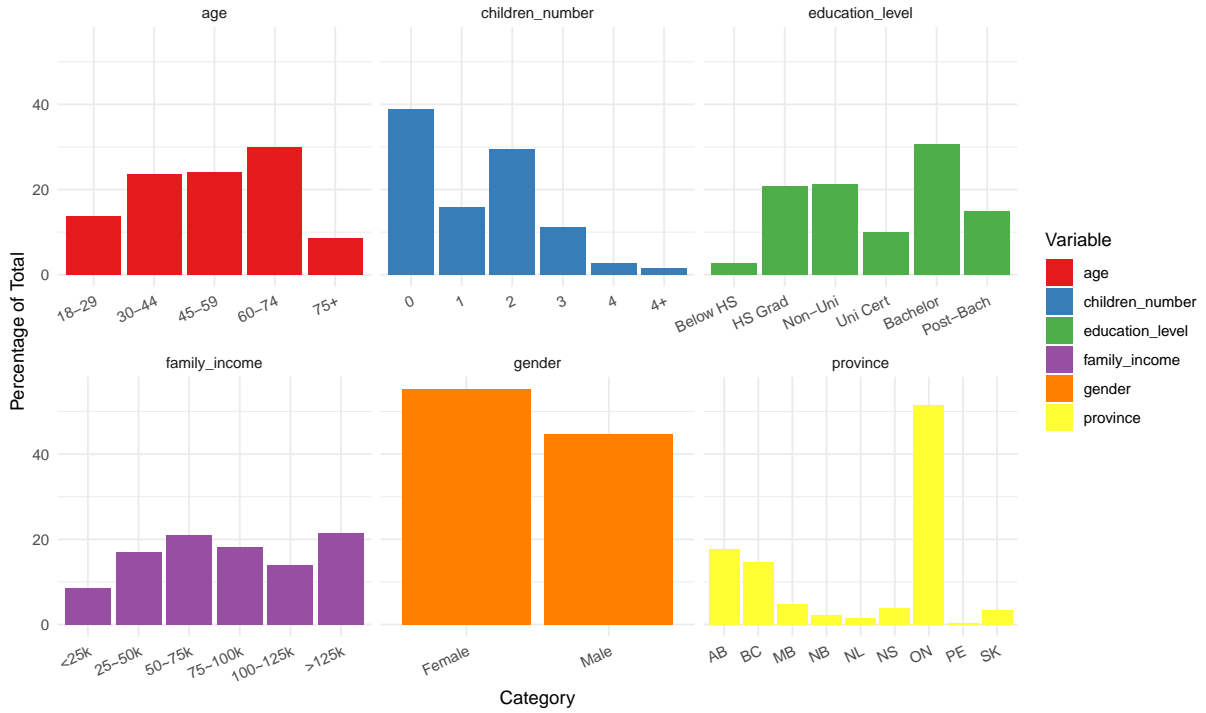


Figure 1: Percentage of different groups in a category

2.5 Measurement

The CES 2021, consisting of an online cohort of 20,968 participants from the broader Canadian populace, was facilitated via the Leger Opinion panel. This study was structured in two distinct stages: the campaign period survey (CPS) and the post-election survey (PES). The

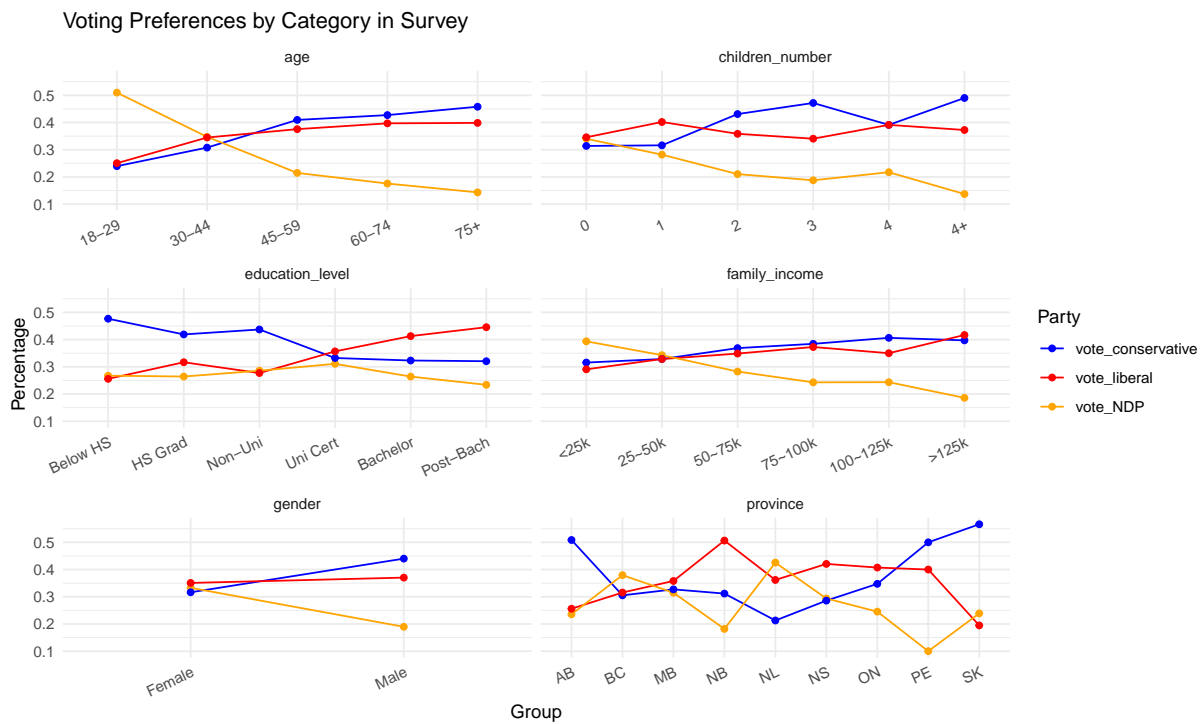


Figure 2: Percentage of Voting for 3 parties in different groups

CPS phase was further divided into three segments – “CPS”, “CPS Modules”, and “CPS Over-sample”, which were later consolidated into one dataset. Regarding the PES, all individuals who participated in the CPS were invited back for this subsequent survey after the election. Out of these, 15,069 individuals completed the PES, marking a response rate of 72%. Respondents needed to be aged 18 or over and Canadian citizens or permanent residents in order to participate.

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 C.

3.1 Multilevel Logistic Regression Model

$$y_i | \pi_i \sim \text{Bern}(\pi_i) \quad (1)$$

$$\text{logit}(\pi_i) = \beta_0 + \alpha_{a[i]}^{\text{age}} + \alpha_{g[i]}^{\text{gender}} + \alpha_{e[i]}^{\text{education level}} + \alpha_{f[i]}^{\text{family income}} + \alpha_{c[i]}^{\text{children number}} + \alpha_{s[i]}^{\text{state}} \quad (2)$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \quad (3)$$

$$\alpha_a^{\text{age}} \sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, 3, 4, 5 \quad (4)$$

$$\alpha_g^{\text{gender}} \sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2 \quad (5)$$

$$\alpha_e^{\text{education level}} \sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, 3, 4, 5, 6 \quad (6)$$

$$\alpha_f^{\text{family income}} \sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, 3, 4, 5, 6 \quad (7)$$

$$\alpha_c^{\text{children number}} \sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, \dots, 6 \quad (8)$$

$$\alpha_s^{\text{state}} \sim \text{Normal}(0, \sigma_{\text{state}}^2) \text{ for } s = 1, 2, \dots, 9 \quad (9)$$

$$\sigma_{\text{state}} \sim \text{Exponential}(1) \quad (10)$$

We run the model in R using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`. The `stan_glm` function from `rstanarm`

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ .

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_1 x_{ij \text{ age}} + \beta_2 x_{ij \text{ gender}} + \beta_3 x_{ij \text{ education level}} + \beta_4 x_{ij \text{ family income}} + \beta_5 x_{ij \text{ children number}}$$

$$\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma_{\beta_0}^2)$$

In the model, j is the province and i is the individual voter indicator.

p represents the probability of the event of interest occurring, which is the voter's probability of voting for any specific party. p_{ij} means the voting of any specific party probability of i th voter in j th province.

$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)$ means the log of odds of voting for any specific party.

β_0 represents the intercept of the model and is the log of odds of voting for any specific party when the individual is at a certain age, with a certain gender, education level, family income and children number.

β_{0j} stands for j th province effect on the log of odds of voting for any specific party, and we assume the intercept β_{0j} in different provinces is random and follows a normal distribution with mean μ_{β_0} and variance $\sigma_{\beta_0}^2$.

β_1 represents the slope of age level in the model, then for any voter, one unit increase in age, we expect a β_1 increase in log odds of voting for any specific party. β_2 represents the average difference in log odds of voting for any specific party between Males and Females for a certain age, education level, family income and children number.

β_3 represents the average difference in log odds of voting for any specific party between groups with different education levels for a certain age, gender, family income and children number.

β_4 represents the average difference in log odds of voting for any specific party between groups with different family income levels for a certain age, gender, education level and children number.

β_5 represents the average difference in log odds of voting for any specific party between groups with different children number levels for a certain age, gender, education level and family income.

x_{ij} is a binary variable, it can only be 1 or 0 as all the variables we have are categorical, it also stands for the i th voter in j th province.

4 Results

4.1 Model Coefficients

5 Discussion

5.1 What is done in this paper?

If my paper were 10 pages, then should be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 What is something that we learn about the world?

5.3 What is another thing that we learn about the world?

5.4 Weaknesses and next steps

Nevertheless, our result's prediction is surprisingly accompanied by recent economic conditions, suggesting a potential winning for the Conservative Party in the upcoming election. Our report relies on the assumption of complete trust in the authenticity of the data used; if responses are recorded in the survey or census showing fake information, it is impossible to eliminate the error caused by that misinformation. Meanwhile, accurate census data depicting the age distribution in the predicted year is essential. The percentage of different ages and the preference for different parties may vary throughout the years, leading to potential inaccuracies in our prediction. The exclusion of Quebec, Yukon, Northwest Territories, and Nunavut from our province variable, either due to distinct political parties within the province (Quebec) or missing data in our sources (Yukon et al.), limits the generalizability of our predictions of these regions. Furthermore, there are different economic conditions between the year of data collected and the year of the prediction. The variation of the economic situation could affect the outcomes of the winning party. Looking forward, analyzing the voting results within Quebec, Yukon, Northwest Territories, and Nunavut satisfies their unique provincial characteristic result better.

Table 2: Coefficients of social factors influence vote for liberal party

	Vote for Liberal Party
(Intercept)	−1.143 (0.262)
age30-44	0.507 (0.139)
age45-59	0.750 (0.143)
age60-74	0.911 (0.147)
age75+	0.933 (0.179)
genderMale	−0.027 (0.078)
education_levelBachelor's degree	−0.028 (0.118)
education_levelHigh school	−0.501 (0.132)
education_levelLess than high school	−0.738 (0.277)
education_levelNon-University	−0.720 (0.130)
education_levelUniversity certificate below the bachelor	−0.188 (0.158)
family_income\$25,000 to \$49,999	0.152 (0.169)
family_income\$50,000 to \$74,999	0.240 (0.162)
family_income\$75,000 to \$99,999	0.308 (0.166)
family_income\$100,000 to \$124,999	0.174 (0.175)
family_income\$125,000 and more	0.385 (0.168)
children_number1	0.088 (0.112)
children_number2	−0.166 (0.101)
children_number3	−0.245 (0.134)
children_number4	−0.075 (0.242)
children_number4+	−0.070 (0.315)
Sigma[province × (Intercept),(Intercept)]	0.187 (0.112)
Num.Obs.	3310
R2	0.060
R2 Marg.	0.042
Log.Lik.	−2066.102
ELPD	−2094.3
ELPD s.e.	20.0

A Appendix

B Additional data details

C Model details

C.1 Posterior predictive check

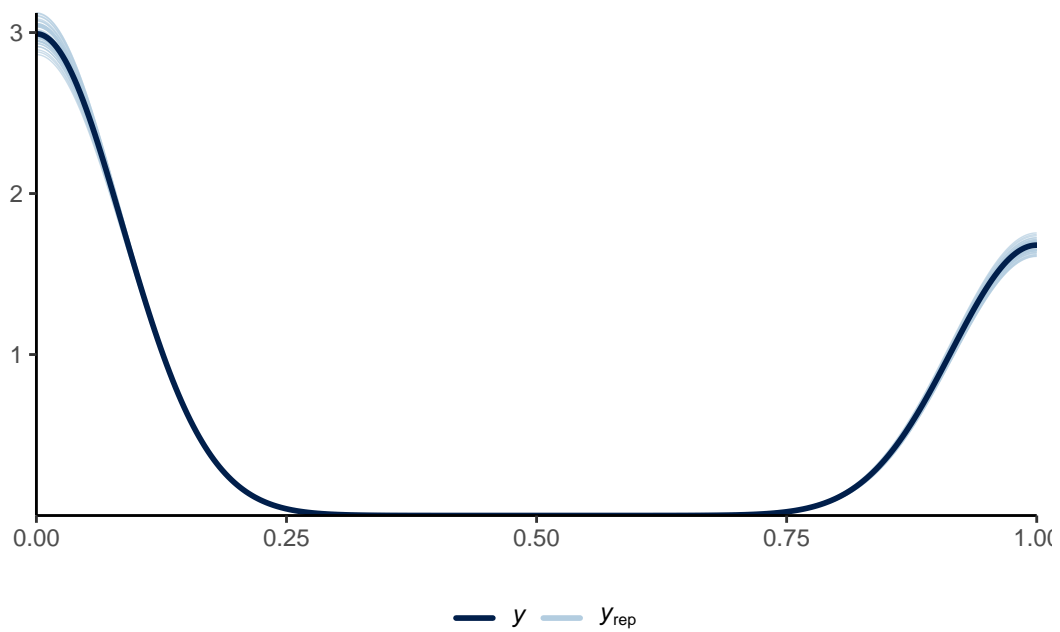


Figure 3: pp check

C.2 Diagnostics

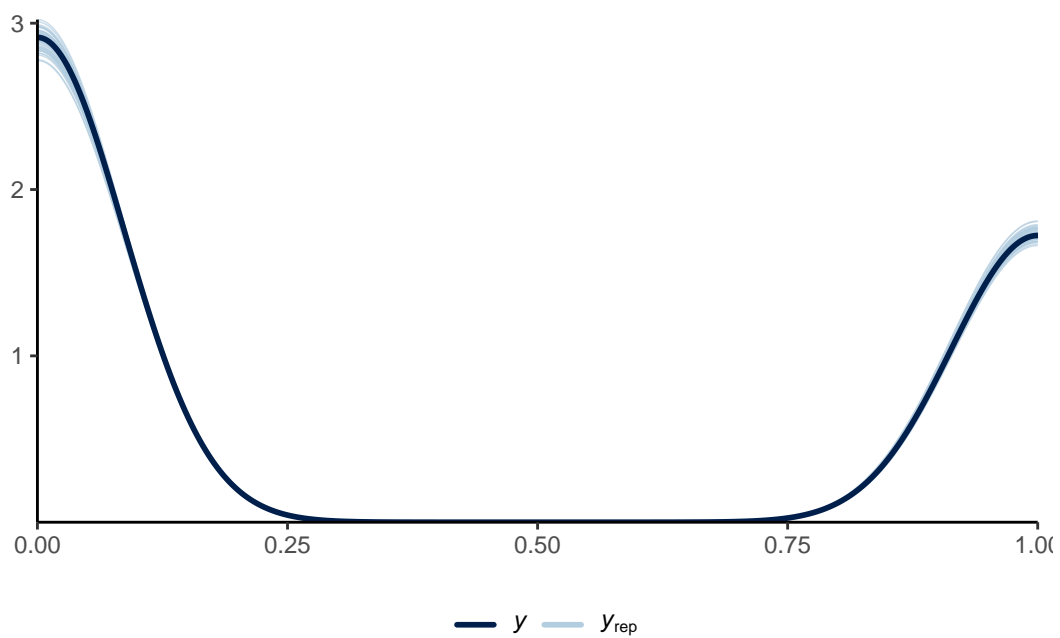


Figure 4: pp check

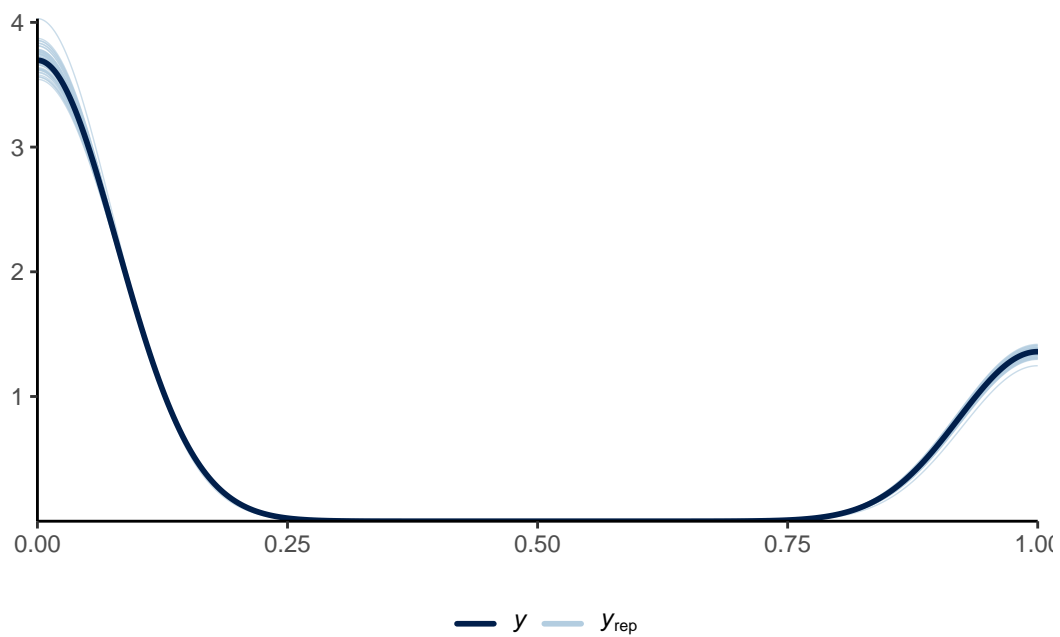


Figure 5: pp check

References

- Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. “Rstanarm: Bayesian Applied Regression Modeling via Stan.” <https://mc-stan.org/rstanarm/>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.