The Impact of Socio-Economic and Demographic Factors on Voter Turnout in Toronto Wards*

An Analysis of the 2022 Toronto Municipal Election

Janel Gilani

April 16, 2024

This paper investigates the impact of socio-economic and demographic factors on voter turnout in the 2022 Toronto Municipal Election. Using data from Open Data Toronto, we use Bayesian analysis to examine the relationships between a ward's voter turnout and factors such as its education level, income, unemployment rate, population, and number of subdivisions. Our inference revealed that while unemployment rate and number of subdivisions have a mild positive and negative correlation with voter turnout respectively, our findings for income, education, and population were not statistically significant. These understandings have implications for attempting to increase voter turnout in future elections.

Table of contents

*Code and data are available at: https://github.com/JanelGilani/toronto-voter-turnout.git

1 Introduction

Voting for the mayor of your ward is one of the most important decisions a Canadian citizen can make because it decides many of the decisions that will be made throughout that four year term as well as how their life will change in that period of time. This is why it is important to understand the many factors that influence the voter turnout and thus, ultimately affect the integrity and representativeness of the electoral process. As mandated by the Government of Ontario, residents of the City of Toronto went to the polls on October 24, 2022 to elect a mayor, councillors, and school board trustees. Then-sitting Mayor, John Tory, sought re-election, along with a number of incumbent city councillors. Coming out of the height of the COVID-19 pandemic, the 2022 election largely upheld the status quo and did not feature ambitious policy platforms. Voter turnout across the city fell to 30% - the lowest in the city's history since amalgamation in 1997, with turnout ranging on a ward-by-ward basis from 22% to 38% (Marshall 2023; Warren 2022).

While there is existing literature on most of these factors in isolation, there has been little research that has examined the combined effect of these explanatory factors on voter turnout, especially in Canadian cities like Toronto and this paper will contribute to the investigation of this phenomenon. Using the 2022 Toronto Municipal Election and Ward Profiles data from Open Data Toronto, this paper will be examining many factors that could have influenced voter turnout in the 2022 Toronto Municipal Election such as education, income, unemployment rate, population, and number of subdivisions within a ward. Our main focus is to conduct analysis on voter turnout as the response variable, and to see if our estimand, the effect of ward's socio-economic and demographic factors on the ward's voter turnout, has influenced voter turnout in the elections. We found that socio-economic factors such as income, unemployment rate, and education level have a significant impact on voter turnout, while demographic factors such as population and number of subdivisions have a less significant impact. Using these understandings, we hope to highlight the potential implications these correlations may have on upcoming elections and the electoral process in Toronto.

Our paper begins with the Data section (Section 2) to visualize and further understand the measurement, source, methodology, and variables we are examining. Then, we introduce the Model (Section 3) used to understand the relationships in the data and report the findings in the Results section (Section 4). Finally, we include the Discussion (Section 5) of the findings, summarizing the takeaway and future of this research.

2 Data

Data analysis is performed in `R` (R Core Team 2022), and additional help is provided by libraries such as `dplyr` (Wickham et al. 2023), `ggplot2` (Wickham 2016), `ggrepel` (Slowikowski 2024), `tidyverse` (Wickham et al. 2019), `kableExtra` (Zhu 2021), `knitr` (Xie 2023), and `sf` (Pebesma and Bivand 2023), `opendatatoronto` (Gelfand 2022), `readxl` (Wickham and Bryan 2023), `here` (Müller 2020), `rstanarm` (Goodrich et al. 2024), `arrow` (Richardson et al. 2023), `tidybayes` (Kay 2023), `modelsummary` (Arel-Bundock 2022), `broom` (Bolker and Robinson 2022), and `parameters` (Lüdecke et al. 2020). Data for this research comes from Open Data Toronto (Gelfand 2022), an open source data portal containing various topics of data for the city. For the data involved in this paper, we combine `Elections - Voter Statistics` (Toronto 2023a) and `Ward Profiles (25-Ward Model)` (Toronto 2023b)

2.1 Measurement

Our research question and estimand is divided into two parts: impact of socio-economic factors and demographic factors on voter turnout, which is our response variable. Starting off with socio-economic factors, we use education, income, and unemployment rate as the explanatory variables to represent these factors. Education is a key factor in determining voter turnout as it is associated with political knowledge, interest, and participation (Blais, Massicotte, and Dobrzynska 2003). To measure education, we use the percentage of citizens with no certificate, diploma, or degree as an indicator of education level. Income and unemployment are also important factor in determining voter turnout as they are associated with financial security, political interest, and participation. Household income is not the sole indicator of wealth, but it is heavily related to wealth as both income and wealth are key indicators of financial security (Schaeffer 2021). We use the unemployment rate as an indicator of economic stability.

Moving on to demographic factors, we use population and number of subdivisions as the explanatory variables to represent these factors. Individual turnout by ward depended on a number of factors, including the accessibility of polling locations within subdivisions to cast a ballot. The City of Toronto Elections (2023) defines a ward as "a geographical area represented by a member of Council." Subdivisions are defined by the City of Toronto (Toronto 2023a) as ".... geographical area[s within a ward] designated by the City Clerk." Previous studies concentrating on large cities in the United States, including Atlanta revealed that having polling locations in close proximity to a voter's home bolsters turnout and even minor changes in placement of a polling location can have significant impact on a voter's decision to cast a ballot (Haspel and Knotts 2005). Thus, we use the number of subdivisions as an indicator of accessibility to polling locations. Lastly, we use population as an indicator of the demographic size of the ward.

Election data is formed by collecting the relevant numbers observed during the phenomenon/process of voting, and ward profiles data is extracted from the 2022 Census data which is collected through a variety of methods such as self-reporting surveys, door-to-door enumeration, online/telephone surveys, administrative records, and sampling techniques.

2.2 Ward Profiles (25-Ward Model)

Ward profiles such as income and population would be interesting factors to analyze alongside eligible voter turnouts. Therefore, the dataset for ward profiles (Toronto 2023b) based on 2022 census data has been included in analysis as well. This dataset is published by City Planning, and was last updated on January 3, 2024. This dataset contains demographic, social, and economic information for each ward such as population, households, families, education, ethnocultural composition, spoken languages, income and housing costs. For our purpose of research, we are interested in population, number of citizens with no certificate, diploma, or degree, unemployment rate, and average household income of each ward. A sample of the cleaned dataset for the wards is shown below in Table 1.

Ward profile data is stored in an Excel file with multiple tabs. The relevant data to be used for this paper's analysis is included the first tab, `2022 Census One Variable`. As such, only data for this tab was downloaded for analysis. Further data cleaning was performed to transpose the data and only keeping information relevant to our research question (see Table 1).

The variables selected for analysis were:

- `Ward ID`: The unique identifier for each ward.
- `Population`: The total number of people living in the ward.

- `Uneducated Population (%)`: The percentage of the population with no certificate, diploma, or degree. This was calculated by dividing the number of uneducated people by the total population and multiplying by 100.
- `Unemployment Rate (%)`: The percentage of the population in labour market that is unemployed.
- `Income`: The average household income in the ward.

Table 1: Sample of Cleaned Toronto Ward Profile Data

Ward ID	Population	Uneducated Population (%)	Unemployment Rate (%)	Income
1	115120	18.997568	16.5	95200
2	117200	11.053754	12.8	146600
3	139920	9.269583	11.8	127200
4	104715	9.072244	12.9	127200
5	115675	21.750594	16.4	88700

2.3 Election Voter Statistics

This dataset, published by the City Clerk's Office (Toronto 2023a) outlines voter statistics on a ward-by-ward and entire city basis for the 2022 municipal election. For each subdivision within a ward, the data set shows the polling location name and address, number of additions and corrections to the voter's list, number of eligible electors and number who voted, and rejected and declined ballots. This data set was last refreshed on February 7, 2023.

The variables (see Table 2) selected for analysis were:

- `Ward ID`: The unique identifier for each ward.
- `Eligible Voter Turnout (%)`: The percentage of eligible voters who cast a ballot.
- `Number of Voters`: The total number of voters who cast a ballot.
- `Number of Subdivisions`: The number of subdivisions within a ward.

Table 2: Sample of Cleaned Elections Data

Ward ID	Eligible Voter Turnout (%)	Number of Voters	Number of Subdivisions
1	24	16962	55
2	30	26784	72
3	33	33544	86
4	38	32223	70
5	28	21807	65

To better visualize the voter turnout across Toronto wards, we have created a bar plot (see Figure 1) that displays the percentage of eligible voters who cast a ballot in each ward. The map of Toronto (see Figure 2) highlights the voter turnout across wards, providing a spatial representation of the data. From Figure 1, we can observe that wards 4, 14, and 19 have the highest voter turnout, while wards 7, 1, and 10 have the lowest voter turnout. This trend confirms the intuition that more affluent and demographically developed areas such as Downtown Toronto have higher voter turnout. The map in Figure 2 further illustrates this trend, with darker shades in southern Toronto indicating higher voter turnout.

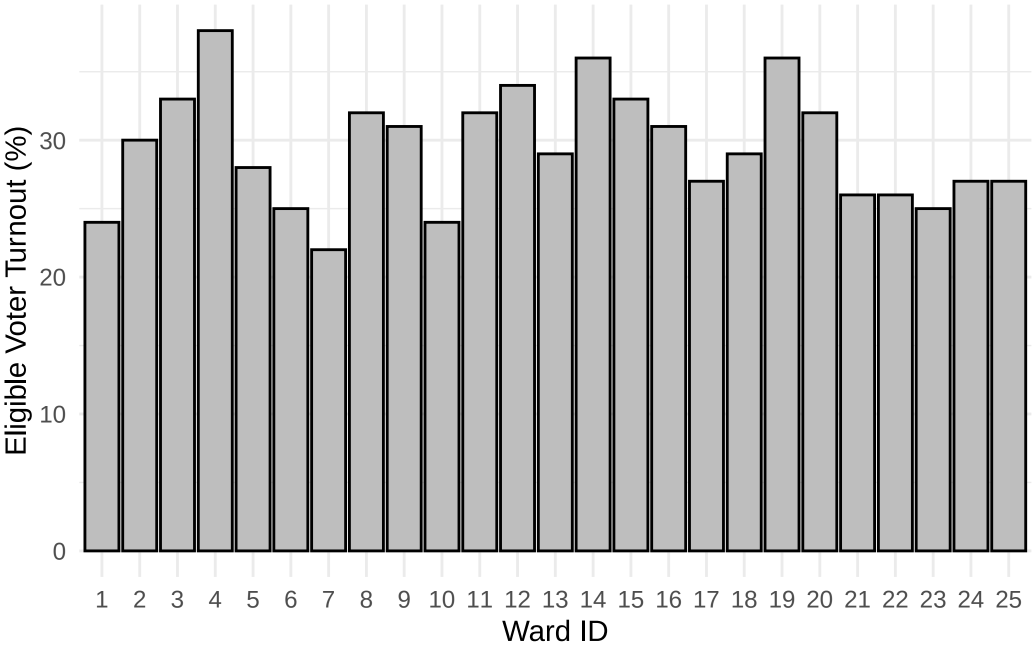


Figure 1: Voter Turnout (%) across Toronto Wards

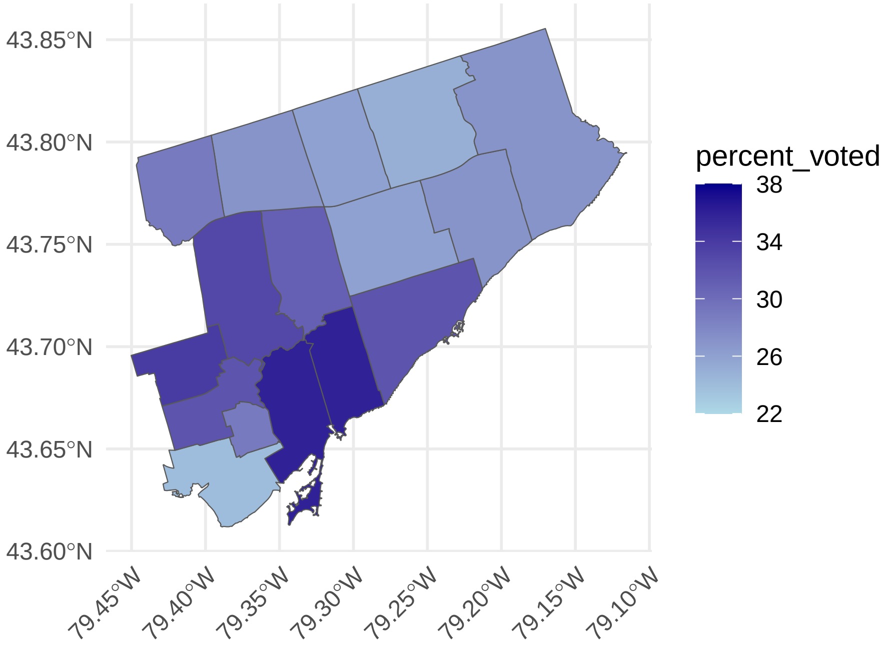


Figure 2: Map of Toronto highlighting the voter turnout across wards

2.4 Voter Turnout and Ward's Socio-Economic and Demographic Factors

As the goal of this research is to analyze the impact of socio-economic and demographic factors on voter turnout in the 2022 Toronto Municipal Election, we have combined the cleaned election data and ward profile data to create an analysis dataset. The analysis dataset includes the following variables: ward ID, ward name, population, number of subdivisions, percent of uneducated population, unemployment rate, income, voter turnout percentage, and number of voters. Below in Table 3 is a sample of the analysis data.

Table 3: Sample of Combined Ward Election, Income, Employment, and Education Data

Ward ID	Name	Pop.	Num. of SubDiv	Uneducated Pop. (%)	Unemployment Rate (%)	Income	Voter Turnout (%)	Num. of Voters
1	Etobicoke North	115120	55	18.997568	16.5	95200	24	16962
2	Etobicoke Centre	117200	72	11.053754	12.8	146600	30	26784
3	Etobicoke-Lakeshore	139920	86	9.269583	11.8	127200	33	33544

Table 3: Sample of Combined Ward Election, Income, Employment, and Education Data

Ward ID	Name	Pop.	Num. of SubDiv	Uneducated Pop. (%)	Unemployment Rate (%)	Income	Voter Turnout (%)	Num. of Voters
4	Parkdale-High Park	104715	70	9.072244	12.9	127200	38	32223
5	York South-Weston	115675	65	21.750594	16.4	88700	28	21807

There are 25 wards in the City of Toronto, each with unique socio-economic and demographic characteristics as summarised in Table 4. Based on 2021 census data, the average population per ward is 110452, with standard deviation of 10594. The wards with the highest population are wards 3, 10 and 2. The wards with the lowest population are wards 23, 16 and 11. The average income per ward is $120096, with standard deviation of $33980.64 The wards with the highest income are wards 15, 8 and 11. The wards with the lowest income are wards 7, 5 and 13.

Moving on to unemployment rate, the average unemployment rate per ward is 14.13%, with standard deviation of 2.11%. The wards with the highest unemployment rate are wards 7, 23 and 1. The wards with the lowest unemployment rate are wards 10, 12 and 3 The average percentage of uneducated population per ward is 12.43%, with standard deviation of 4.93%. The wards with the highest percentage of uneducated population are wards 7, 5 and 1. The wards with the lowest percentage of uneducated population are wards 10, 11 and 13.

Voter turnout across the entire city during the 2022 municipal election fell to 30%, with Ward 4 (Parkdale-High Park) having the highest turnout at 38% and Ward 7 (Humber River-Black Creek) having the lowest at 22%. Ward 19 (Beaches-East York) and Ward 14 (Toronto Danforth) were tied for second highest turnout at 36% each. The third highest turnout at 34% was in Ward 12 (Toronto-St. Paul's). The wards with the second lowest turnout were Ward 1 (Etobicoke North) and Ward 10 (Spadina-Fork York) at 24% each. Ward 6 (York Centre) and Ward 23 (Scarborough North) with 25% turnout respectively were tied for third lowest turnout.

During the 2022 municipal election, there were 1,535 subdivisions across Toronto's 25 wards. Ward 10 (Spadina-Fort York) had the highest number of subdivisions at 94, while Ward 23 (Scarborough North) had the lowest number at 38. Ward 13 (Toronto Centre) had the second highest number of subdivisions at 90, with Ward 3 (Etobicoke-Lakeshore) coming in third highest with 83 subdivisions. Ward 7 (Humber River-Black Creek) had the second lowest number of subdivisions at 47, following by Ward 25 (Scarborough-Rouge Park) with 48.

Table 4: Summary Statistics

Variable	Mean	Median	Std. Dev	Min	Max
Income	120096.00	107300.00	33980.64	85700.00	224800.00
Number of Subdivisions	61.40	59.00	13.79	38.00	94.00
Number of Voters	22524.96	22522.00	5168.51	14616.00	33544.00
Population	110451.60	110095.00	10593.87	94025.00	139920.00
Voter Turnout (%)	29.48	29.00	4.22	22.00	38.00
Uneducated Population (%)	12.43	11.95	4.93	4.84	23.08
Unemployment Rate (%)	14.13	14.10	2.11	9.80	17.80

2.5 Relationship between Voter Turnout and Socio-Economic Factors

As one of our variables of interest, we are determined to examine the relationship between voter turnout and the ward's income. We expect income and voter turnout to be positively related because Higher-income individuals often have better access to resources, education, and stability, which can positively correlate with higher levels of political engagement and voter turnout. To visualize the relationship of interest, we plot ward's income with the voter turnout in Figure 3.

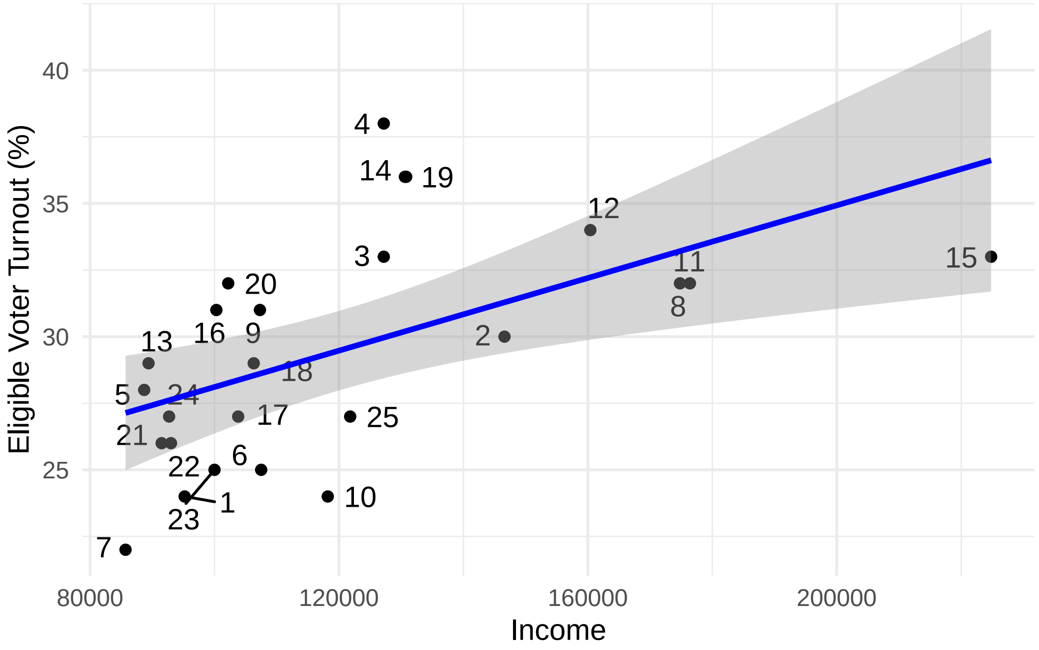


Figure 3: Correlation between Eligibile Voter Turnout and Ward's Income

As expected, we see a moderate positive relationship between income of a ward and the voter turnout. Furthermore, Ward 4 is an outlier that has relatively low income and yet a high voter turnout. Intuitively, this aligns with our beliefs and confirms the trend between high income with higher voter turnout.

Another explanatory variable we are interested in is the employment rate of a region. In this case, we visualize the relationship between employment rate and the nvoter turnout per ward. We expect to see a negative relationship between these variables. Figure 4 displays the visualization.

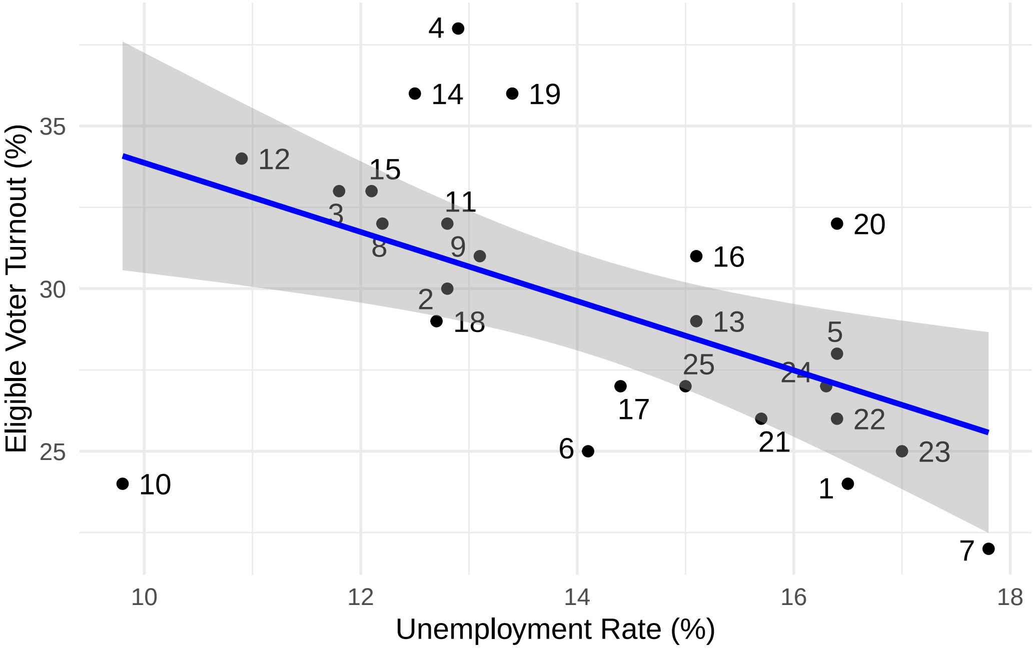


Figure 4: Correlation between Eligibile Voter Turnout and Ward's Unemployment Rate

Through the plot, we see that there is a moderately negative relationship between the unemployment rate of a ward and the voter turnout. This is in line with our expectations as higher levels of employment often correlate with increased stability and engagement within a community. Employed individuals may feel more connected to their local area and thus be more likely to participate in civic activities such as voting. Ward 10 is an outlier that has a low unemployment rate and yet a low voter turnout. This observation leaves room for further research and investigation into the history and background behind Ward 10's voter turnout.

Lastly, we examine the relationship between the percentage of uneducated population in a ward and the voter turnout. We expect to see a negative relationship between these variables. Figure 5 displays the visualization.

The plot in Figure 5 shows a negative relationship between the percentage of uneducated

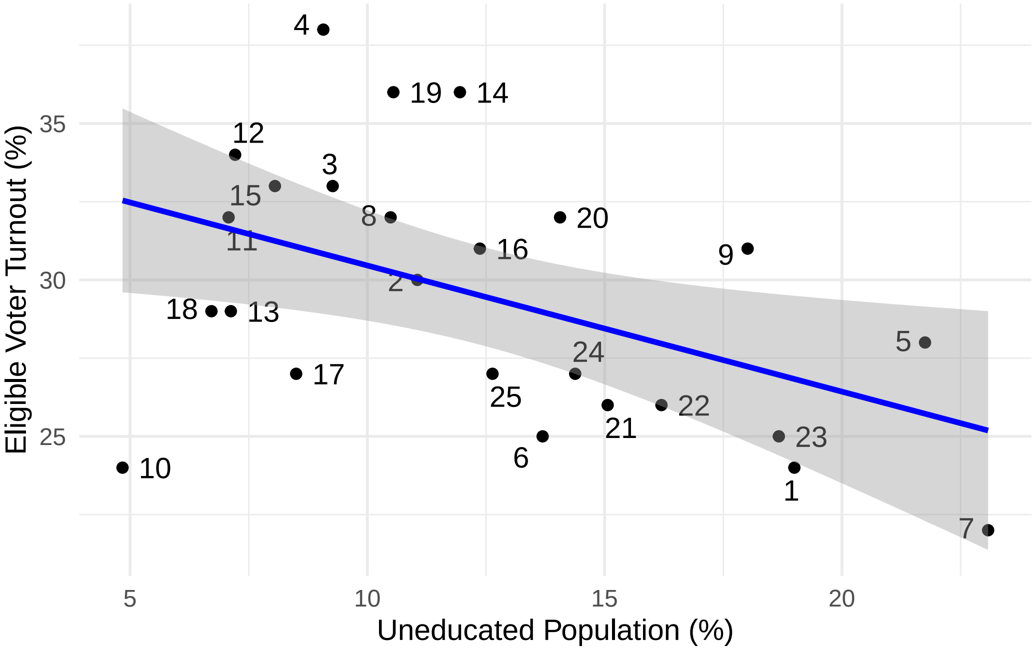


Figure 5: Correlation between Eligibile Voter Turnout and Ward's Level of Education

population in a ward and the voter turnout. This also aligns with our expectations as higher education levels typically correlate with greater awareness of civic duties, including voting, and a stronger sense of civic responsibility, which can lead to higher voter turnout. Ward 10 is again an outlier that has a low percentage of uneducated population and yet a low voter turnout.

2.6 Relationship between Voter Turnout and Demographic Factors

In addition to socio-economic factors, we are also interested in examining the relationship between voter turnout and demographic factors such as population and number of subdivisions. We expect to see a positive relationship between population and voter turnout as higher population density in a ward may indicate greater community engagement and political activity, leading to higher voter turnout. Additionally, densely populated areas often have more resources and infrastructure, making it easier for residents to access polling stations and participate in elections. To visualize this relationship, we plot the ward's population with the voter turnout in Figure 6.

Through the plot, we see that there is no relationship between the population of a ward and the voter turnout. It is possible that the population of a ward is not a reflection of its full demographic nature. Thus, we would not need to worry about population acting as a confounding factor when creating our demographic models.

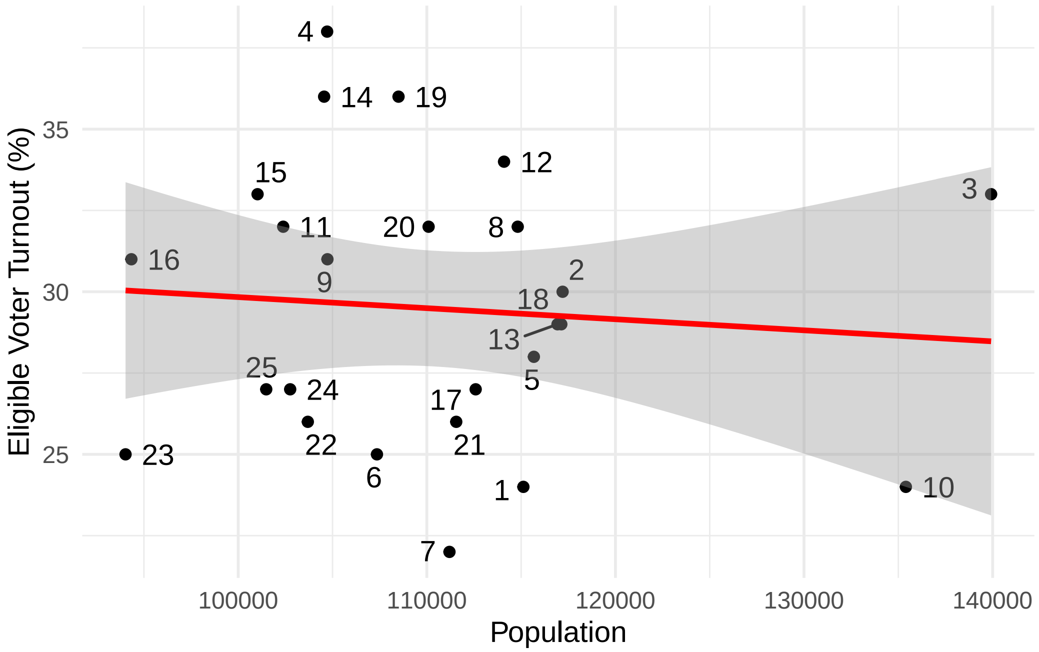


Figure 6: Correlation between Eligibile Voter Turnout and Ward's Population

Next, we examine the relationship between the number of subdivisions in a ward and the voter turnout. We expect to see a positive relationship between these variables as a higher number of subdivisions may indicate greater accessibility to polling stations and increased voter turnout. Figure 7 displays the visualization.

Even though we see a slight positive correlation, the data reveals that having more subdivisions within wards does not automatically correlate to higher voter turnout, as illustrated by Figure 7. Ward 4 (Parkdale-High Park) which had the highest voter turnout at 38% has 70 subdivisions within the ward, which falls approximately in the middle of the number of subdivisions across all 25 wards. Ward 19 (Beaches-East York) and Ward 14 (Toronto Danforth) which both saw the second highest voter turnout at 36% had 57 and 52 subdivisions respectively, once again highlighting that more subdivisions does not necessarily correlate to higher voter turnout.

3 Model

Here we briefly describe the Bayesian analysis model used to investigate the relationship between the socio-economic and demographic factors and voter turnout. In particular, we divide our investigation into two parts: socio-economic models and demographic models. For each, we consider two types of models: multiple linear regression and Poisson regression models. We

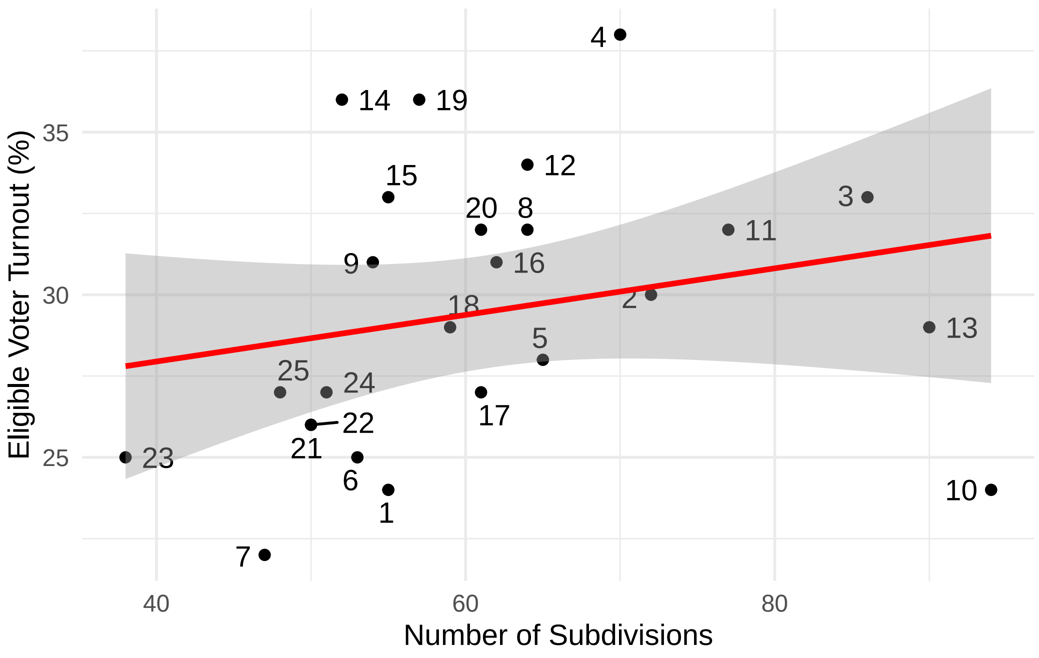


Figure 7: Correlation between Eligibile Voter Turnout and Ward's Number of Subdivisions

also further tested our results from Poission regression by employing negative binomial regression models to account for unequal mean and variance in the data (The negative binomial regression results are not displayed here but the models can be found in relevant files in the `models` folder).

Background details and diagnostics are included in Appendix B.

3.1 Model Set-Up

From the Data section, we observed the relationships between voter turnout and socio-economic and demographic factors in isolation. Now, we aim to understand the relationships in a more comprehensive manner by adjusting for all factors simultaneously using the following models. We create and run the generalised linear models in R (R Core Team 2022) using the `rstanarm` package of Goodrich et al. (2024). Initially, we use the default priors from `rstanarm`, however, we allow `rstanarm` to improve the priors by scaling them based on the data. We allow auto-scaling and run both models with the updated priors specified above.

3.1.1 Socio-Economic Models

Table 5 shows a model summary for socio-economic models, this will be discussed further in the results section.

3.1.1.1 Poisson Regression

Define y_i as the number of voters in the ward i. Then $income_i$ is the income of ward i, $uneducated_i$ is the uneducated population percentage of ward i, and $unemployment_i$ is the unemployment rate of ward i.

$$\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (1)\\
log(\lambda_i) &= \beta_0 + \beta_1 \times \text{uneducated}_i + \beta_2 \times \text{income}_i + \beta_3 \times \text{unemployment}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 0.508) && (4)\\
\beta_2 &\sim \text{Normal}(0, 0.074) && (5)\\
\beta_3 &\sim \text{Normal}(0, 1.187) && (6)
\end{aligned}$$

3.1.1.2 Multiple Linear Regression

Define y_i as the voter turnout percentage in the ward i. Then $income_i$ is the income of ward i, $uneducated_i$ is the uneducated population percentage of ward i, and $unemployment_i$ is the unemployment rate of ward i.

$$\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (7)\\
\mu_i &= \beta_0 + \beta_1 \times \text{uneducated}_i + \beta_2 \times \text{income}_i + \beta_3 \times \text{unemployment}_i && (8)\\
\beta_0 &\sim \text{Normal}(0, 11) && (9)\\
\beta_1 &\sim \text{Normal}(0, 2.14) && (10)\\
\beta_2 &\sim \text{Normal}(0, 0.31) && (11)\\
\beta_3 &\sim \text{Normal}(0, 5.01) && (12)\\
\sigma &\sim \text{Exponential}(0.24) && (13)
\end{aligned}$$

3.1.2 Demographic Models

Table 6 shows a model summary for demographic models, this will be discussed further in the results section.

3.1.2.1 Poisson Regression

Define y_i as the number of voters in the ward i. Then $population_i$ is the population of ward i, and $numsub_i$ is the number of subdivisions in ward i.

$$\begin{aligned} y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (14)\\ log(\lambda_i) &= \beta_0 + \beta_1 \times \text{numsub}_i + \beta_2 \times \text{population}_i && (15)\\ \beta_0 &\sim \text{Normal}(0, 2.5) && (16)\\ \beta_1 &\sim \text{Normal}(0, 0.18125) && (17)\\ \beta_2 &\sim \text{Normal}(0, 0.00024) && (18) \end{aligned}$$

3.1.2.2 Multiple Linear Regression

Define y_i as the voter turnout percentage in the ward i. Then $population_i$ is the population of ward i, and $numsub_i$ is the number of subdivisions in ward i.

$$\begin{aligned} y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (19)\\ \mu_i &= \beta_0 + \beta_1 \times \text{numsub}_i + \beta_2 \times \text{population}_i && (20)\\ \beta_0 &\sim \text{Normal}(0, 11) && (21)\\ \beta_1 &\sim \text{Normal}(0, 0.766) && (22)\\ \beta_2 &\sim \text{Normal}(0, 0.0001) && (23)\\ \sigma &\sim \text{Exponential}(0.24) && (24) \end{aligned}$$

3.2 Model Justification

We chose to use Bayesian analysis to investigate the relationship between socio-economic and demographic factors and voter turnout because it allows us to incorporate prior knowledge and uncertainty into our models. By using Bayesian analysis, we can estimate the posterior distribution of the parameters of interest, which provides a more comprehensive understanding of the relationships between the variables. Additionally, Bayesian analysis allows us to quantify the confidence levels in the estimates of our coefficients and make probabilistic statements about the relationships between the variables.

For our first model, we used a Poisson regression model to investigate the relationship between voter turnout and socio-economic factors. We chose this model because it is appropriate for count data, such as the number of voters, and allows us to model the relationship between the explanatory variables and the response variable. We also chose to use a multiple linear regression model to investigate the relationship between voter turnout and socio-economic factors. We included income, unemployment rate, and the percentage of uneducated population as explanatory variables in the model. We chose this model because it allows us to examine the relationships between multiple variables and voter turnout simultaneously.

We then repeated the analysis for demographic factors, using population and the number of subdivisions as explanatory variables. We used the same models as for the socio-economic factors to investigate the relationships between these variables and voter turnout. We chose these models because they allow us to examine the relationships between the variables and voter turnout while controlling for other factors.

From Appendix Figure 10a, we see that the Poisson regression model is not a good fit for the observed data. To improve our model, we consider the multiple linear regression model with all the explanatory variables except using percentage of voter turnout instead of unnormalised voter count. From Appendix Figure 10b, we see that the multiple linear regression is an improved fit from the Poisson regression model. This could be because the key assumption that the mean and variance are equal is violated. From Table 4, we see that mean and variance are not equal.

Since an important assumption for the Poisson regression model does not hold, we also build a negative binomial model. We can relax the assumption of mean and variance as equal in negative binomial model. We got almost identical results from the negative binomial regression model as it is a close variant of the Poisson model with looser assumptions whose details can be found in the relevant files in the `models` folder.

4 Results

4.1 Socio-Economic Models

Table 5 and Figure 8 shows the results of the socio-economic models.

In the multiple linear regression model, the coefficients represent the change in the response variable (voter turnout) for a one-unit increase in the predictor variable, holding all other variables constant. In this analysis, the coefficient for the percentage of the uneducated population is -0.11, suggesting that for every one-unit increase in the percentage of uneducated individuals within a ward, the voter turnout decreases by 0.11 percentage points, although this effect is not statistically significant at the 99% confidence level. Conversely, the coefficient for income is 0.04, indicating that for every one-thousand increase in income, the voter turnout increases by 0.04 percentage points. However, this relationship is also not statistically significant. The coefficient for the unemployment rate is -0.41, implying that for every one-unit increase in the unemployment rate, the voter turnout decreases by 0.41 percentage points. While this coefficient is negative, indicating a decrease in voter turnout with higher unemployment rates, the relationship is not statistically significant either. The intercept term represents the expected voter turnout when all predictor variables are zero, which is 32.47% compared to the mean value of 29.48%. While this may never be practical in real life, it tells us that the negative correlation of voter turnout with variables like unemployment rate and uneducated population is relatively stronger than the positive correlation with income.

In the Poisson regression model, the coefficients represent the log-linear relationship between the predictor variables and the expected count of the response variable (number of voters). For the percentage of the uneducated population, the coefficient is -0.08, indicating that a one-unit increase in the percentage of uneducated individuals within a ward is associated with an 8% decrease in the expected count of voters, though this effect is not statistically significant at the 99% confidence level. The coefficient for income is 0.00, which implies that changes in income do not significantly affect the expected count of voters in the Poisson model. Similarly, the coefficient for the unemployment rate is -0.08, suggesting that increases in unemployment may lead to a slight decrease in the expected count of voters, although this effect is also not statistically significant. Overall, the Poisson regression results suggest that the percentage of uneducated individuals within a ward may have a modest impact on the expected count of voters, while income and unemployment rate do not significantly influence voter turnout in terms of the number of voters.

These coefficient values may be influenced by various factors related to how predictor variables and voter turnout are related. For instance, the negative coefficient for the unemployment rate suggests that higher unemployment rates may contribute to lower voter turnout. This could be due to factors such as decreased civic engagement among individuals facing economic hardship or logistical barriers to voting for those experiencing unemployment. However, the lack of statistical significance may indicate that other unmeasured variables or complex interactions between factors play a more significant role in determining voter turnout. Additionally, the relatively small coefficient values for income and the percentage of uneducated individuals suggest that these variables may have weaker effects on voter turnout within the context of the studied population. Overall, while the coefficients provide insight into the direction and potential magnitude of relationships between predictor variables and voter turnout, the lack of statistical significance suggests that these relationships may not be robust or generalizable across all wards in Toronto.

4.2 Demographic Models

Table 6 and Figure 9 shows the results of the demographic models.

In the multiple linear regression model, the coefficients represent the estimated change in the response variable (voter turnout) associated with a one-unit increase in the predictor variables, while holding all other variables constant. For the intercept, the coefficient is 39.80, indicating the estimated voter turnout when all predictor variables are zero. This value is statistically significant at the 99% confidence level, with a confidence interval of [12.40, 63.82]. The coefficient for the number of subdivisions is 0.17, suggesting that for each additional

Table 5: Socio-economic explanatory models of voter turnout based on education, employment and income

	Multiple Linear Regression	Poisson Regression
(Intercept)	31.28	11.03
	[9.90, 52.02]	[11.00, 11.06]
percent_uneducated	−0.11	0.00
	[−0.61, 0.39]	[0.00, 0.00]
income	0.04	0.00
	[−0.02, 0.10]	[0.00, 0.00]
unemployment_rate	−0.41	−0.08
	[−1.79, 0.97]	[−0.08, −0.07]
Num.Obs.	25	25
R2	0.354	
R2 Adj.	0.105	
Log.Lik.	−66.902	−7176.104
ELPD	−71.8	−7687.8
ELPD s.e.	4.8	1719.9
LOOIC	143.7	15 375.7
LOOIC s.e.	9.6	3439.7
WAIC	142.4	18 010.8
RMSE	3.33	3685.45

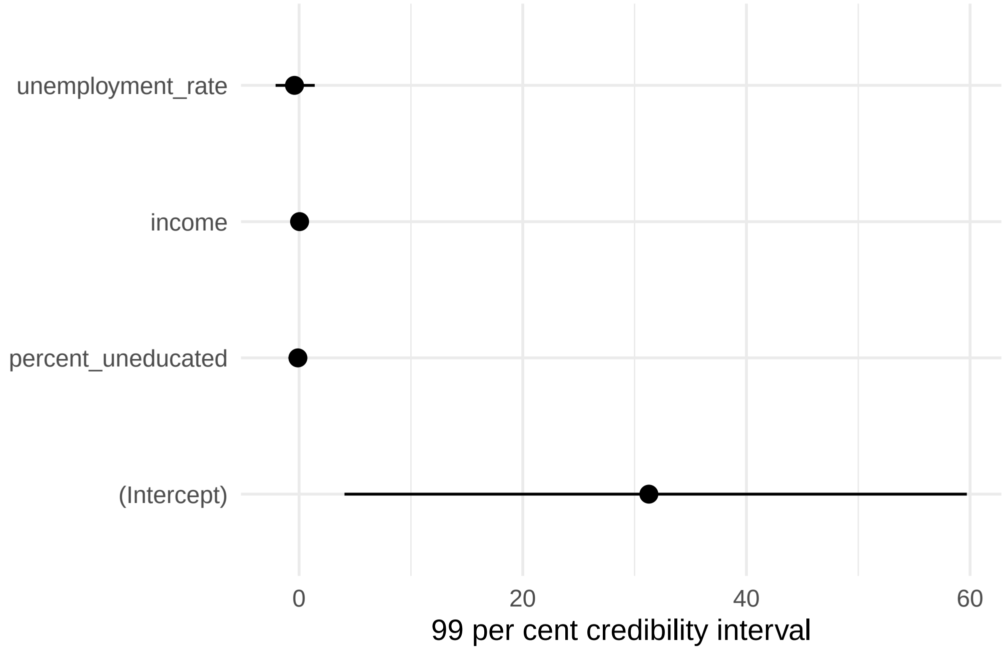


Figure 8: Explanatory model of voter turnout based on socio-economic factors

subdivision within a ward, the expected voter turnout increases by 0.17 units. However, this effect is not statistically significant at the 99% confidence level, as the confidence interval [-0.08, 0.39] includes zero.Similarly, the coefficient for population is 0.00, indicating that changes in population size do not significantly influence voter turnout in terms of the number of voters. The confidence interval [0.00, 0.00] further supports this finding, suggesting that the effect size of population on voter turnout is negligible.

Overall, while the intercept provides an estimate of baseline voter turnout, the coefficients for the number of subdivisions and population do not demonstrate significant associations with voter turnout in the multiple linear regression model. However, it's important to consider the limitations of the model and potential confounding factors that may impact these relationships.

In the Poisson regression model, the coefficients represent the log of the expected count of the response variable (number of voters) associated with a one-unit increase in the predictor variables, while holding all other variables constant. For the intercept, the coefficient is 9.52, indicating the estimated log count of voters when all predictor variables are zero. This value is statistically significant at the 99% confidence level, with a confidence interval of [9.48, 9.56]. The coefficient for the number of subdivisions is 0.01, suggesting that for each additional subdivision within a ward, the expected log count of voters increases by 0.01 units. This effect is statistically significant at the 99% confidence level, as the confidence interval [0.01, 0.01] does not include zero. Similarly, the coefficient for population is also 0.00, indicating that

Table 6: Demographic explanatory models of voter turnout based on population and number of subdivisions

	Multiple Linear Regression	Poisson Regression
(Intercept)	39.80	9.52
	[12.40, 63.82]	[9.48, 9.56]
num_sub	0.17	0.01
	[−0.08, 0.39]	[0.01, 0.01]
population	0.00	0.00
	[0.00, 0.00]	[0.00, 0.00]
Num.Obs.	25	25
R2	0.185	
R2 Adj.	−0.106	
Log.Lik.	−69.574	−9817.594
ELPD	−73.3	−10 373.2
ELPD s.e.	3.0	1962.9
LOOIC	146.5	20 746.4
LOOIC s.e.	6.0	3925.8
WAIC	146.1	23 902.6
RMSE	3.77	4268.57

changes in population size do not significantly influence the expected log count of voters. The confidence interval [0.00, 0.00] further supports this finding.

Overall, the Poisson regression model suggests that the number of subdivisions within a ward has a statistically significant, although very small, positive association with the expected count of voters. However, changes in population size do not have a significant impact on the expected count of voters. It's important to interpret these findings in the context of the Poisson distribution and the assumptions of the model.

5 Discussion

5.1 Key Findings about Socio-Economic Factors

In this paper, we aimed to verify our expectations that socio-economic factors such as income, unemployment rate, and the percentage of uneducated individuals within a ward would im-

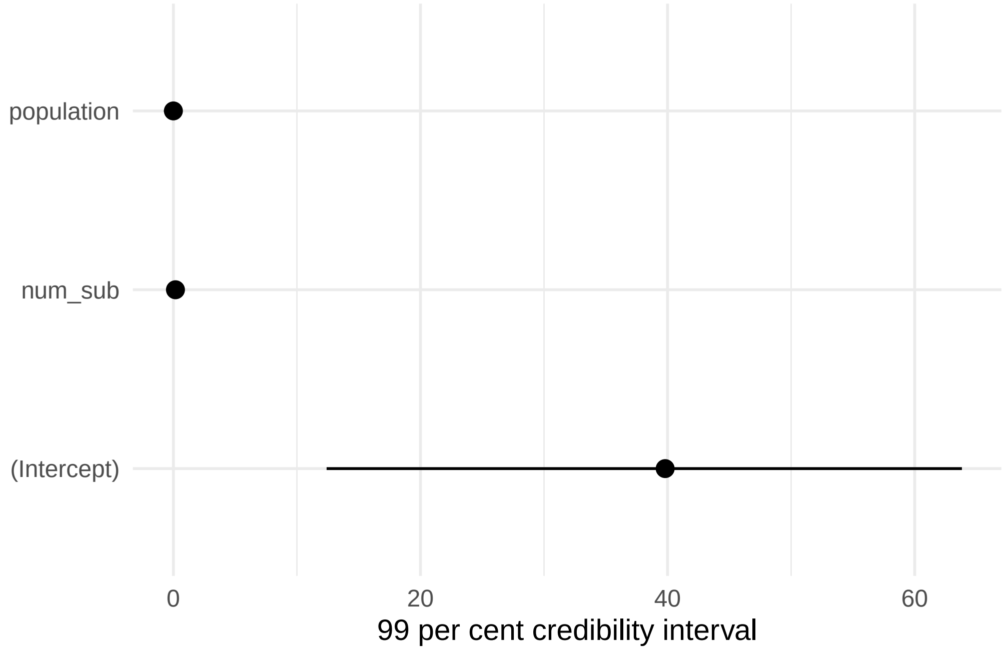


Figure 9: Explanatory model of voter turnout based on demographic factors

pact voter turnout in the 2022 Toronto municipal election. During exploratory analysis of data, we observed that while income was positively correlated with voter turnout, the unemployment rate and the percentage of uneducated individuals were negatively correlated with voter turnout. This led us to hypothesize that higher income levels would be associated with higher voter turnout, while higher unemployment rates and a higher percentage of uneducated individuals would be associated with lower voter turnout. This hypothesis is backed by existing literature, as individuals with higher incomes often have greater access to resources, education, and stability, which can facilitate political engagement and higher voter turnout. Conversely, higher unemployment rates and a higher percentage of uneducated individuals may indicate economic insecurity and limited access to information or civic participation opportunities, contributing to lower voter turnout among these groups.

However, our subsequent Bayesian analysis Table 5 revealed that the relationships between these socio-economic factors and voter turnout were not statistically significant at the 99% confidence level. While the coefficients in the models suggested that higher income levels were associated with higher voter turnout, and higher unemployment rates and a higher percentage of uneducated individuals were associated with lower voter turnout, the lack of statistical significance indicates that these relationships may not be robust or generalizable across all wards in Toronto. This suggests that the relationships between socio-economic factors and voter turnout may be more complex and nuanced than initially anticipated, and that other unmeasured variables or confounding interactions between factors may play a more significant role in determining voter turnout.

These findings can be utilised by policymakers and community leaders to develop targeted strategies and interventions to increase voter turnout in Toronto. For example, efforts to increase voter turnout among lower-income or less educated populations could focus on providing accessible information about the election process, offering resources to facilitate voting, and addressing barriers to political participation. Similarly, initiatives to engage unemployed individuals in the electoral process could involve outreach programs, community partnerships, and targeted messaging to encourage voter registration and turnout. By understanding the socioeconomic factors that influence voter turnout, stakeholders can develop tailored approaches to promote civic engagement and increase voter participation in future elections.

5.2 Key Findings about Demographic Factors

In terms of demographic factors, we expected that population density and the number of subdivisions within a ward would impact voter turnout in the 2022 Toronto municipal election. We hypothesized that higher population density and a higher number of subdivisions would be associated with higher voter turnout, as these factors may indicate greater community engagement, accessibility to polling stations, and resources that facilitate political participation. However, our exploratory analysis revealed that while population density was not correlated with voter turnout, the number of subdivisions had a weak positive correlation with voter turnout. This led us to further investigate the relationships between these demographic factors and voter turnout using Bayesian analysis.

The demographic models in Table 6 showed that the number of subdivisions within a ward had a statistically significant, although very small, positive association with the expected count of voters in the Poisson regression model. This suggests that an increase in the number of subdivisions within a ward may lead to a slight increase in the expected count of voters. However, changes in population size did not significantly influence the expected count of voters in either the multiple linear regression or Poisson regression models. These findings align with our initial hypothesis that the number of subdivisions may impact voter turnout, as subdivisions may provide greater accessibility to polling stations and facilitate community engagement. However, our analysis could not conclusively establish a relationship between population and voter turnout.

These results can inform strategies to increase voter turnout in Toronto by highlighting the importance of local infrastructure and community resources in promoting civic engagement. Efforts to increase voter turnout could focus on enhancing accessibility to polling stations, providing information about voting procedures, and fostering community connections within subdivisions. By understanding the demographic factors that influence voter turnout, policymakers and community leaders can develop targeted interventions to promote voter participation and strengthen democratic engagement in Toronto.

5.3 Weaknesses and Next Steps

From the Appendix posterior prediction checks in Figure 10, and Figure 11, we can see that the Poisson regression models do not fit the data well but multiple linear regression does fairly well. This is likely because our data violates an important assumption of the Poisson regression model: equal mean and variance. In the data, we have 25 observations for the variables since we are comparing data at the ward level across the 25 Toronto wards. The small number of observations combined with the model fit indicates a potential data problem in the research. The relatively small sample size of 25 observations is a critical factor to consider, as it might not only impact the robustness of the statistical models but also reflect on the generalizability of the findings. Small sample sizes can lead to higher variability and may affect the model's ability to accurately capture the underlying distribution of the data.

Moreover, the mismatch between the expected model conditions and the observed data characteristics suggests that there may be underlying factors affecting the data quality or distribution that were not accounted for in the initial analysis. This could range from measurement errors to unaccounted-for variables that could significantly influence the outcomes of the ward-level comparisons. Therefore, this analysis does not merely highlight a statistical anomaly but points to a larger data problem that could have implications for the research's validity and reliability.

The cause of not being able to establish statistically significant relationships between most socio-economic and demographic factors and voter turnout could be due to the limitations of the dataset. The lack of detailed information on the specific characteristics of each ward, such as the presence of community organizations, local initiatives, or historical voting patterns, may have limited the depth of the analysis. Future research could benefit from incorporating additional data sources or conducting qualitative research to gain a more comprehensive understanding of the factors influencing voter turnout in Toronto. Additionally, expanding the dataset to include more observations or variables could help address the limitations of the current analysis and provide a more robust foundation for investigating the relationships between socio-economic and demographic factors and voter turnout.

Next, the creation of subdivision boundaries within wards lacks transparency and clarity, while impacting the local democratic process. The City Clerk is responsible for drawing up all subdivisions, however the data and information they use to make these decisions is not public and it is challenging to account for any bias introduced during this process. Moreover, the number of subdivisions and population alone cannot dictate voter behaviour, as there are many other demographic factors which inform why an elector does or does not cast a ballot. The caliber of local candidates and the nature of the election, whether an incumbent is standing for re-election or the seat is open for a new councillor impacts voter turnout and should be accounted for in future analysis.

As mentioned previously, our small data problem can affect the internal validity of the research in question. Since we have a small data set, there is a reduction in statistical power, the probability of correctly rejecting a false null hypothesis. Additionally, having a small number of observations increases the risk of overfitting. In this case, the model learns the noise in the data instead of the underlying pattern. The data has limited number of observations and is at the ward level in Toronto, so generalizing our findings to cities outside Toronto may raise external validity concerns. Toronto's unique socio-economic, cultural, and environmental characteristics may influence the study's variables in ways that are not replicable in other cities. Factors such as policies and economic conditions vary significantly from one city to another, potentially affecting the applicability of the findings elsewhere.

With the purpose of investigating the relationship between socioeconomic and demographic factors on voter turnout in Toronto, we find a moderate positive relationship between number of subdivisions and voter turnout in ward and a slight negaitive correlation between unemployment rate and voter turnout supported by data visualization as well as statistical models. The research is unable to conclude any relationship between income and the percentage of uneducated individuals with voter turnout. The research is also unable to conclude any relationship between population and voter turnout. Violation in the Poisson model assumptions as well as a small data set calls for a larger data set in future research to address both internal and external validity concerns. With a larger number of observations, model inferences will improve, addressing internal validity. To expand the data set, we can consider gathering data from other cities. Through this, we are also able to address external validity. With the inclusion of a large data set with many cities, we can gather data with socio-economic diversity. When working with more diverse data, we can potentially generalize our findings to elections across the world, helping us learn more about the factors that affect voter turnout in a variety of contexts.

Appendix

A Datasheet

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset was created to provide voter statistics for the 2022 municipal election in Toronto. It includes information on the number of eligible electors, additions and corrections to the voter's list, and the number of voters who cast ballots, broken down by ward and subdivision.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The dataset was created by the City Clerk's Office on behalf of the City of Toronto.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The creation of the dataset was funded by the City of Toronto.

4. *Any other comments?*

    - No comments.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The instances in the dataset represent the wards of the City of Toronto. Each instance corresponds to a specific ward within the city, providing voter statistics for the 2022 municipal election.

2. *How many instances are there in total (of each type, if appropriate)?*

    - There are 25 instances in total, each representing a unique ward within the City of Toronto.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset contains all possible instances of wards within the City of Toronto for the 2022 municipal election. As such, it is not a sample but rather a complete enumeration of all wards and their corresponding voter statistics for the specified election period. Therefore, the dataset is representative of the entire city and does not require validation or verification of representativeness.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of voter statistics for a specific ward within the City of Toronto for the 2022 municipal election. The dataset includes 23 continuous variables, such as the number of eligible electors, the number of voters, and the eligible voter turnout percentage. Additionally, each instance includes a categorical feature identifying the ward represented.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Yes, the label associated with each instance is the ward ID, which represents the unique identifier for each ward within the City of Toronto.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - There is no missing information from individual instances. The dataset provides comprehensive voter statistics for each ward, and all relevant features are included without any missing values.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - There are no relationships between individual instances.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - There are no recommended data splits.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - There are no errors, sources of noise, or redundancies in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - There is no confidential data, and the dataset is publicly available.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No, the dataset does not contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The dataset does not identify any sub-populations such as age or gender. It primarily focuses on voter statistics for different wards in the City of Toronto, without specific demographic information about individual voters. Therefore, there are no identifiable subpopulations within the dataset.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - It is not possible to identify individuals in any way.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

- None.

16. *Any other comments?*

    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data associated with each instance was acquired through interviews conducted with subjects from all 25 states. Subjects reported the data directly, providing information relevant to voter statistics for the respective wards in the City of Toronto.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The mechanism used to collect the data was manual human curation. This involved individuals conducting interviews and recording the information provided by the subjects. The validation of this process likely involved quality checks during data entry and potentially cross-referencing the collected information with other sources to ensure accuracy and consistency.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset represents all possible instances, meaning it contains data for every ward in the City of Toronto. Therefore, there was no sampling strategy involved as the dataset encompasses the entire population of interest. However, if there were sampling involved, the strategy would likely have been geographic stratification, subdividing each state into regions and systematically selecting primary sampling units with probability proportional to the size of the strata.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data collection process likely involved individuals from the City Clerk's Office or other relevant departments within the municipal government of Toronto. These individuals would have been responsible for conducting interviews, gathering voter statistics, and maintaining the dataset. It's reasonable to assume that they were compensated as part of their regular employment with the municipal government, although specific details regarding compensation are not provided in the dataset information.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- The data was last refreshed on February 7, 2023, as indicated in the dataset details. This timeframe corresponds to the last update or revision of the dataset. The data collection process likely occurred during the 2022 municipal election period, which took place on October 24, 2022. Therefore, the data collection timeframe aligns with the creation timeframe of the data associated with the instances.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - Ethical review processes were not conducted.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - We obtained the data via the Open Data Toronto website

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - The dataset information does not specify whether individuals were notified about the data collection process. However, given that the data is publicly available and pertains to voter statistics for the 2022 municipal election, it is likely that individuals were aware of the data collection process through public records and disclosures related to the election.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - The dataset information does not provide details on how consent was obtained from individuals for data collection and use. Given that the data pertains to voter statistics for the 2022 municipal election, it is likely that consent was not explicitly requested or required, as this information is typically considered public record and collected as part of the election process.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - A mechanism to revoke consent was not provided.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - There is no mention of an analysis of the potential impact of the dataset and its use on data subjects.

12. *Any other comments?*

    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - We cleaned the data by removing any unnecessary columns and renaming the columns to make them more descriptive. We also converted the data types of certain columns to ensure consistency and ease of analysis.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - The raw data downloaded using Open Data Toronto is available in the folder data/raw_data/raw_data.csv.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - R Software is avalaible at https://www.R-project.org/

4. *Any other comments?*

    - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The dataset has not been used for other tasks yet.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - No, there is no repository that links to any papers or systems that use the dataset.

3. *What (other) tasks could the dataset be used for?*

    - The election dataset could be used for various analyses, such as predicting voter turnout, identifying trends in voter participation, and evaluating the impact of socio-economic factors on voting behavior. Additionally, the dataset could be used to assess the effectiveness of voter registration campaigns, analyze the distribution of voters across different wards, and explore the relationship between voter turnout and demographic characteristics.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - The dataset contains voter statistics for the 2022 municipal election in Toronto and does not include any personal information about individual voters. However, dataset consumers should be aware that the data pertains to voter participation and may be subject to privacy and confidentiality considerations. To avoid potential risks or harms, consumers should use the data responsibly and ensure that any analyses or interpretations are conducted in an ethical and unbiased manner. Additionally, consumers should be mindful of the limitations of the dataset and consider the context in which the data was collected when using it for future tasks.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - Not applicable.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - No, the dataset is openly available and being used for personal uses only.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- The dataset is distributed using Open Data Toronto and will be available for download as a CSV file. The cleaned dataset is available on Github.

3. *When will the dataset be distributed?*

    - The raw dataset is already available for download from Open Data Toronto. The cleaned dataset will be available on Github on 18th April 2024.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The dataset has been released under the Open Government License - Toronto. The cleaned dataset used in this paper will be available on Github under the MIT License.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - There are no restrictions

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No such controls or restrictions are applicable.

7. *Any other comments?*

- None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - Janel Gilani

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - Can be contacted via github

3. *Is there an erratum? If so, please provide a link or other access point.*

    - There is no erratum available currently.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - Currently there is no plan of updating the dataset. If there are updates in the future, it will be done through Github.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - The dataset was made publicly available and does not contain personal information about individuals. Therefore, there are no limits on the retention of data associated with the instances. The data will be retained indefinitely for research and analysis purposes.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - The older versions would not be hosted. Dataset consumers will be able to check whether the dataset has been updated through Github commit history.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - There is no mechanism for accepting contributions from other users as of now.

B Model Details

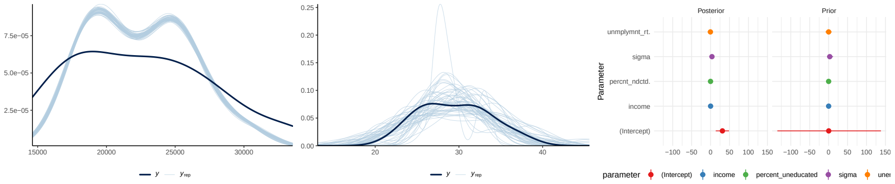
B.1 Posterior Predictive Check

In Figure 10a we implement a posterior predictive check on the Poisson regression model. This shows the comparison between the actual outcome variable (number of voters) with simulations from the posterior distribution. From the figure, we can see that the observed data has peaked while the posterior predictive distributions are more dispersed. This means that the model is not a good fit and does not replicate the observed distribution well.

In Figure 10b we implement a posterior predictive check on the multiple linear regression model. This shows the comparison between the actual outcome variable (voter turnout percentage) with simulations from the posterior distribution. Here the observed data and posterior predictions have some overlap. This model is a better fit than the multiple regression model.

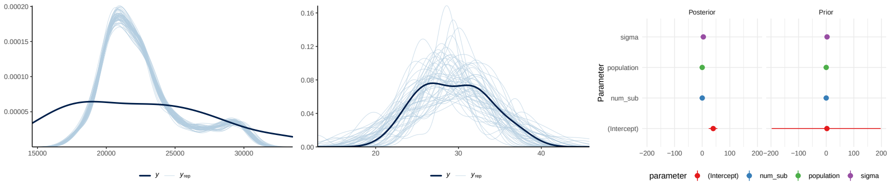
In Figure 10c we compare the posterior with the prior. This shows how much the estimates of the coefficients of our variables: unemployment rate, uneducated population and income, have changed once data was taken into account.

Similar results are seen for the demographic models in Figure 11a, Figure 11b, and Figure 11c.



(a) Poisson (b) Multiple Linear (c) Posterior v/s Prior

Figure 10: Examining how the socio-economic models fit, and is affected by, the data



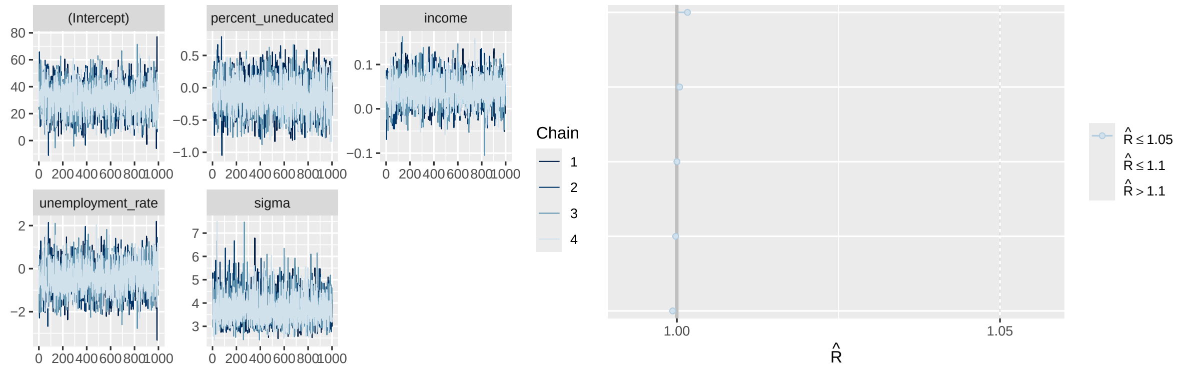
(a) Poisson (b) Multiple Linear (c) Posterior v/s Prior

Figure 11: Examining how the demographic models fit, and is affected by, the data

B.2 Diagnostics

The Markov chain Monte Carlo sampling algorithm checks for signs that the algorithm has issues. We consider a trace plot and a Rhat plot for our socio-economic model. In Figure 12a, we see horizontal lines that bounce around and have overlap between the chains. In Figure 12b, we see that everything is close to 1. We do not see anything out of the ordinary in the trace plot or Rhat plot, indicating that the algorithm did not run into any issues.

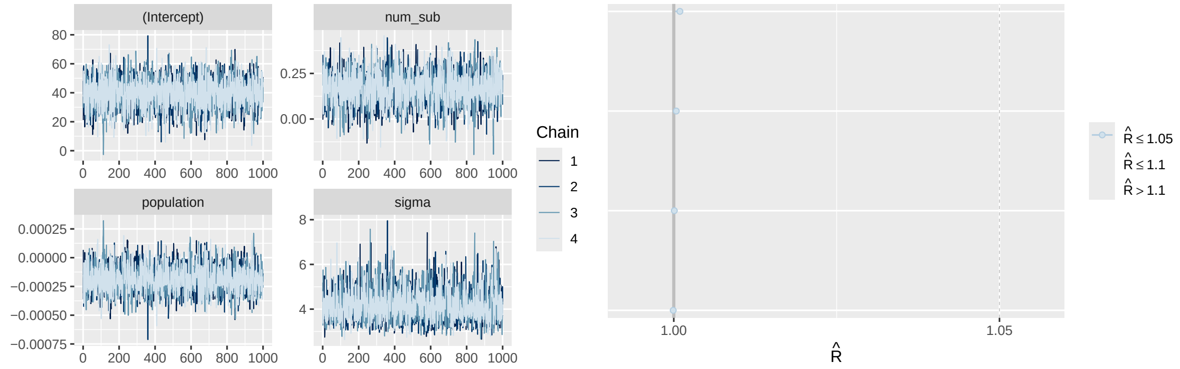
We see the same results for the demographic model in Figure 13a and Figure 13b.



(a) Trace plot

(b) Rhat plot

Figure 12: Checking the convergence of the MCMC algorithm for the socio-economic multiple linear regression model



(a) Trace plot

(b) Rhat plot

Figure 13: Checking the convergence of the MCMC algorithm for the demographic multiple linear regression model

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.

Blais, Andre, Louis Massicotte, and Agnieszka Dobrzynska. 2003. "Why Is Turnout Higher in Some Countries Than in Others?" https://www.elections.ca/res/rec/part/tuh/TurnoutHigher.pdf.

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Elections, Toronto. 2023. "Election Dictionary." https://www.toronto.ca/city-government/elections/city-elections/education-resources/election-dictionary/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of ACM* 64 (12): 86–92.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://sharlagelfand.github.io/opendatatoronto/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Haspel, Moshe, and H Gibbs Knotts. 2005. "Location, Location, Location: Precinct Placement and the Costs of Voting." *The Journal of Politics* 67 (2): 560–73.

Kay, Matthew. 2023. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, and Dominique Makowski. 2020. "Extracting, Computing and Exploring the Parameters of Statistical Models Using R." *Journal of Open Source Software* 5 (53): 2445. https://doi.org/10.21105/joss.02445.

Marshall, Sean. 2023. "Election: Voter Turnout 2022." http://spacing.ca/toronto/2023/03/07/election-voter-turnout-in-2022/.

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

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With Applications in r*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429459016.

R Core Team. 2022. *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. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Schaeffer, Katherine. 2021. "'What's the Difference Between Income and Wealth?' And Other Common Questions about Economic Concepts." *Decoded*. Pew Research Center. https://www.pewresearch.org/decoded/2021/07/23/whats-the-difference-between-income-and-wealth-and-other-common-questions-about-economic-concepts/.

Slowikowski, Kamil. 2024. *Ggrepel: Automatically Position Non-Overlapping Text Labels with*

'Ggplot2'. https://ggrepel.slowkow.com/.

Toronto, Open Data. 2023a. "Elections - Voter Statistics." https://open.toronto.ca/dataset/elections-voter-statistics/.

———. 2023b. "Ward Profiles (25-Ward Model)." https://open.toronto.ca/dataset/ward-profiles-25-ward-model/.

Warren, May. 2022. "Toronto 2022 Municipal Election Brings Dismal Voter Turnout." https://www.thestar.com/news/gta/2022/10/24/toronto-2022-municipal-election-brings-dismal-voter-turnout.html.

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, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. https://readxl.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

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*. http://haozhu233.github.io/kableExtra/.