Determinants of Wage Variation in Canada*

Older Workers, Men, and Higher Education Earn Higher Average Hourly Wages

Yuanyi (Leo) Liu

December 2, 2024

This paper examines wage disparities in Canada by analyzing the effects of education, gender, and age on average hourly earnings. The findings indicate that older workers consistently earn higher wages, men earn more than women across all groups, and higher levels of education are strongly associated with increased wages. These patterns highlight systemic differences in earnings tied to demographic and educational factors. Understanding these disparities provides important context for policymakers aiming to address wage inequality and improve economic outcomes.

Table of contents

*Code and data are available at: Determinants of Wage Variation in Canada.

1 Introduction

Wages are a fundamental component of economic well-being, influencing individual livelihoods, workforce productivity, and broader societal equity. Understanding the factors that shape earnings is essential for addressing wage inequality and improving economic opportunities. In Canada, wage disparities persist across demographic and educational groups, raising important questions about the relationship between age, gender, and education and their roles in determining hourly wages.

This paper analyzes how average hourly wages in Canada vary based on three key factors: age, gender, and education level. Using data spanning two decades, we employ a linear regression model to quantify the relationships between these factors and earnings. The analysis examines whether older workers earn higher wages, how gender influences pay, and the extent to which education impacts earnings. These questions are important in the ongoing concerns about income inequality and the need for policies that promote equitable pay structures. While previous studies have documented broad patterns of wage inequality, this paper offers a focused examination of specific demographic and educational factors, providing a deeper understanding of their combined effects.

In this study, our estimand is the average hourly wage rate in Canada. The object of the estimation is the average hourly wage rate across different demographic and educational groups based on the data. By modeling these relationships, the study quantifies the extent to which

these factors contribute to observed wage disparities, providing a detailed understanding of their relative influence within the Canadian labor market.

The findings demonstrate that wages increase with age, particularly for workers over 55, while men consistently earn more than women across all groups. Higher levels of education are strongly associated with higher wages, with university graduates earning substantially more than those with lower educational attainment. These results highlight persistent disparities that reflect both structural and systemic influences on earnings. By examining these relationships, this paper provides a basis for understanding wage inequality in Canada and informs discussions about potential policy interventions.

The structure of the paper is organized as follows: following Section 1, Section 2 presents the data collection and cleaning process, along with an overview of the variables used in the analysis. Section 3 explains the chosen model and why it is appropriate for modeling average hourly wages. Then, Section 4 provides the results, highlighting key trends and predictions. Eventually, Section 5 concludes with a discussion of the findings, addressing wage disparities, potential strategies to improve equity in Canadian earnings, and the limitations of the model used.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to process and analyze data on average hourly wages in Canada. The dataset (Statistics Canada 2020), published on January 11, 2020, covers data from January 1, 1997, to December 31, 2019, but for this analysis, we focus on data from 2000 onwards to ensure relevance to recent labor market trends. The dataset is maintained annually and provides detailed information on wages, demographics, and employment characteristics, making it well-suited for studying wage disparities in Canada. Following methodologies discussed in "Telling Stories with Data" (Alexander 2023), we analyze wage patterns by aggregating data across multiple demographic groups to ensure a balanced and unbiased representation of wage disparities. For key operations, please refer to Appendix B.

The dataset includes variables such as age groups (15+, 25+, 25-34, 25-54, and 25-64), types of work (full-time and part-time), educational levels (e.g., 0-8 years, high school graduate, post-secondary certificate diploma, university degree), and immigration statuses (e.g., very recent immigrants, recent immigrants, established immigrants, non-landed immigrants, and Canadian-born individuals). Wages are provided as both average weekly and average hourly rates. For this analysis, we focus on the age groups "15-24 years", "25-54 years", and "55 years and over", the combined category of full- and part-time workers, and specific education levels, such as high school graduate, post-secondary certificate diploma, and university degree.

This filtering excludes ambiguous or aggregate categories such as "some high school," "some post-secondary," and "total landed immigrants" to ensure a clear and interpretable dataset. This filtering allows us to analyze relationships between wages and specific demographic factors while maintaining focus on education, gender, and age. Alternative datasets, such as those focusing on industry-specific wage trends or detailed time-series data, were not used because they do not provide the necessary demographic granularity. This dataset is uniquely positioned to address our research questions by capturing a broad view of wage disparities in Canada.

2.2 Measurement

Wage data is a measurement of economic outcomes based on information collected through the Labour Force Survey Special Tabulations conducted by Statistics Canada. Respondents' demographic information, including age, gender, education level, and employment type, is reported through standardized surveys administered to a representative sample of the Canadian workforce. Wage data, including both hourly and weekly rates, are derived from employer records and worker-reported earnings, ensuring broad coverage across different industries and occupations. More details on survey methodologies can be found in the Labour Force Survey Special Tabulations.

Each row in the dataset represents an aggregated statistic for a specific combination of age group, education level, gender, and type of work. For example, the average hourly wage rate for "25-54 years" old male workers with a "university degree" is calculated by pooling the wages of individuals who fit this description, smoothing out individual variations to provide a general trend for this subgroup. By categorizing variables such as education levels into distinct groups (e.g., "high school graduate," "post-secondary certificate diploma"), the dataset allows for consistent comparisons across demographic segments.

2.3 Outcome variables

2.3.1 Hourly Wages

The primary outcome variable in this study is the `Average_hourly_wages`, measured in Canadian dollars. This variable reflects the mean earnings of individuals across various demographic and employment categories. The hourly wage data are derived from the Labour Force Survey, which collects self-reported income information and supplements it with employer records where available. These averages are calculated by dividing reported weekly wages by the number of hours worked, standardizing earnings across full-time and part-time workers.

Table 1: Summary Statistics for Average Hourly Wages in Canada (2000–2019), Including Minimum, Maximum, Mean, Median, and Standard Deviation

Minimum Wage	Maximum Wage	Mean Wage	Median Wage	Standard Deviation
7.6	43.39	21.48	20.48	7.4

Table 1 provides an overview of the distribution of average hourly wages in Canada from 2000 to 2019. The wages range from a minimum of $7.60 to a maximum of $43.39, with a mean of $21.48 and a median of $20.48, indicating that most wages are concentrated around the middle of the distribution, while a smaller proportion of higher earners pushes the maximum upward. The standard deviation of $7.40 reflects notable variability, emphasizing differences in earning potential influenced by factors such as education, gender, and age.

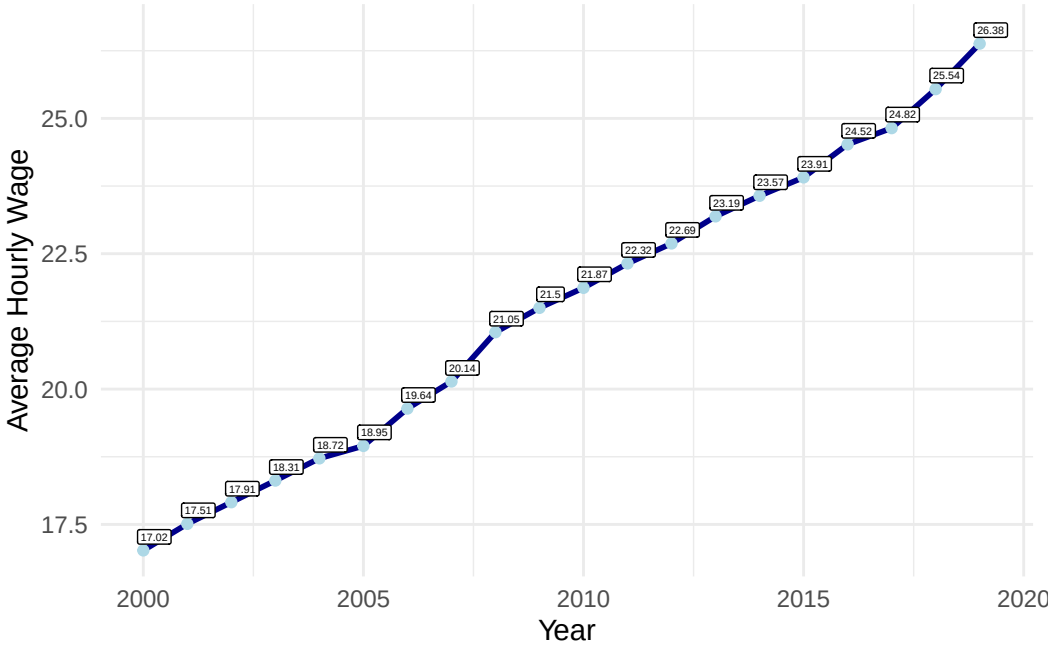

Figure 1: Yearly Trend of Average Hourly Wages in Canada (2000–2019): Steady Growth Reflecting Changes in the Labor Market.

Figure 1 illustrates the trend of average hourly wages in Canada from 2000 to 2019. Each dot represents the average wage for a specific year, and the overall shows a upward trend. The consistent increase in wages suggests improvements in earnings across the labor market, potentially driven by inflation, economic growth, and changes in workforce composition.

2.4 Predictor variables

2.4.1 Education

The `Education_level` variable represents individuals' highest level of formal education attained and is categorized into levels ranging from "0-8 years" to "Above bachelor's degree." This variable reflects the influence of educational attainment on earning potential. The data for education levels are self-reported in the Labour Force Survey, capturing a wide range of qualifications, including high school completion, trade certifications, and university degrees.

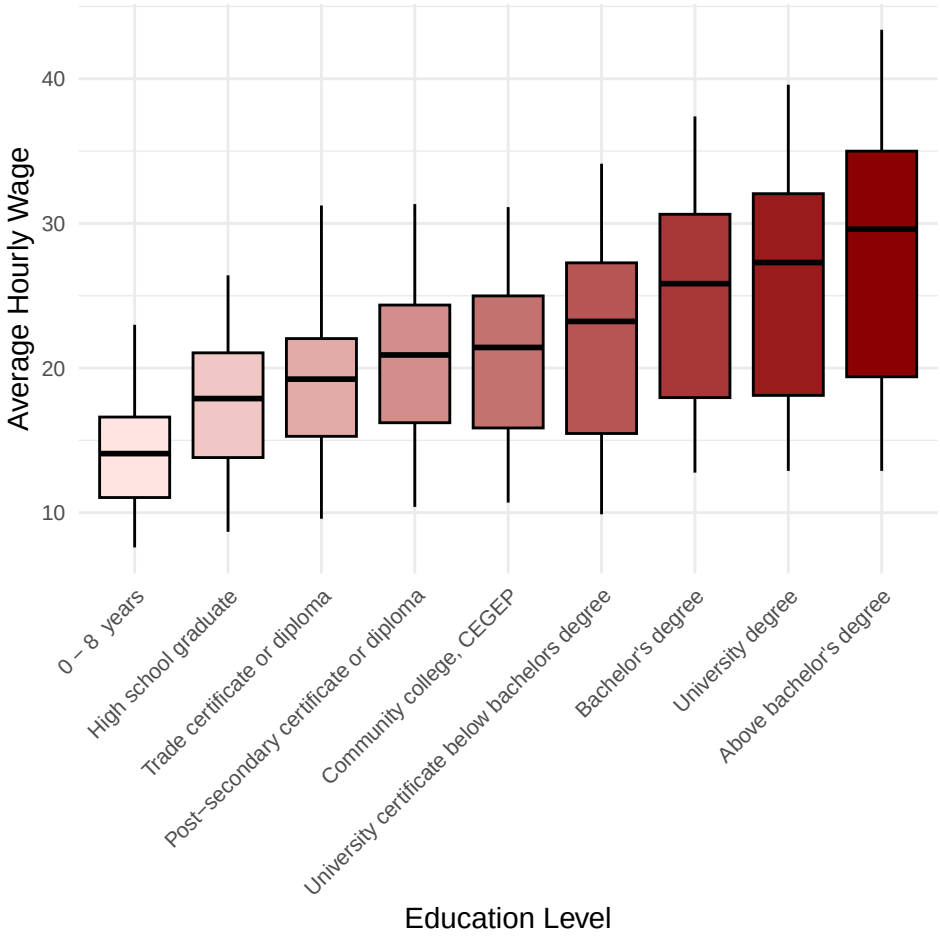

Figure 2: Average hourly wages in Canada by education level from 2000 to 2019. Each box represents the wage distribution for a specific education level, with higher education levels corresponding to higher median wages.

Figure 2 presents the distribution of average hourly wages in Canada from 2000 to 2019 across different education levels. Higher education levels, such as "University degree" and "Above

bachelor's degree," are associated with higher median wages, while lower education levels, like "0–8 years," show significantly lower wages. The variability within each education level, shown by the spread of the boxes and whiskers, indicates differences in earning potential within these groups. For instance, individuals with an "Above bachelor's degree" earn higher wages on average and exhibit greater wage variability compared to those with "0–8 years" of education.

2.4.2 Gender

The `Gender` variable indicates whether an individual identifies as "Male" or "Female". The data are reported as binary classification for simplicity. This variable allows for an investigation of the persistent gender wage gap, capturing differences in average hourly wages for men and women. Including gender in the analysis is essential for highlighting inequalities in the labor market and understanding how these disparities persist across educational levels and age groups.

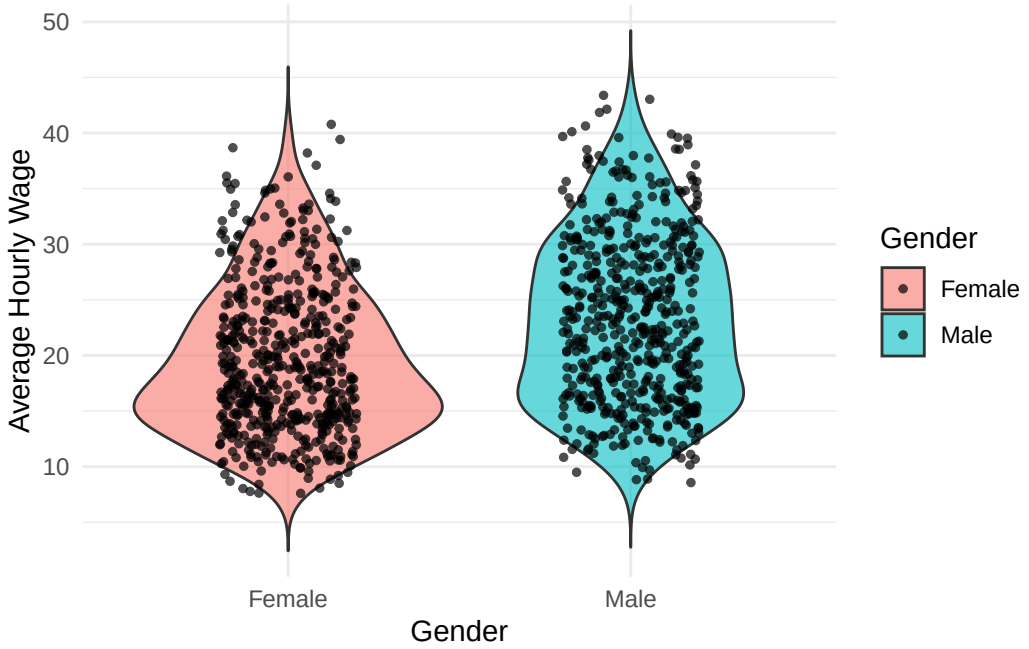

Figure 3: Average hourly wages in Canada by gender from 2000 to 2019. Each violin plot represents the wage distribution for males and females, with males showing higher median wages and a broader range of earnings compared to females.

Figure 3 compares the distribution of average hourly wages in Canada from 2000 to 2019 by gender. It displays the spread of wages for each gender, with males generally earning higher wages as shown by the broader and higher distribution. The jittered points represent individual wage observations, further emphasizing the overlap and differences in wage ranges.

While both distributions have a similar shape, the median wage for males is noticeably higher than that for females, reflecting a persistent gender wage gap over the period analyzed.

2.4.3 Age

The `Age_group` variable divides individuals into three categories: "15-24 years", "25-54 years", and "55 years and over". These groups represent different career stages, from early employment to peak earning years and late career. The Labour Force Survey provides these classifications to standardize analyses of how wages vary with experience and age. Younger workers generally report lower wages due to limited experience, while mid-career and older workers often earn higher wages, reflecting accumulated skills and knowledge over time.

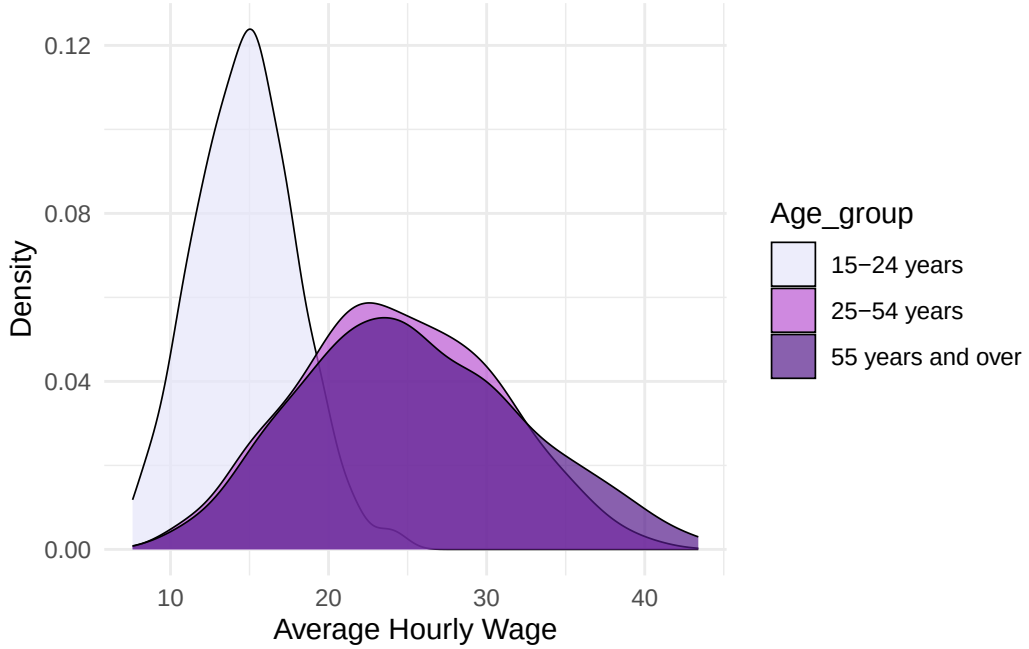

Figure 4: Average hourly wages in Canada by age group from 2000 to 2019. Each curve represents the wage distribution for a specific age group, with older age groups showing higher average wages and broader distributions compared to younger groups.

Figure 4 illustrates the distribution of average hourly wages in Canada from 2000 to 2019 across three age groups: "15-24 years," "25-54 years," and "55 years and over." The youngest group, "15-24 years," exhibits the lowest wages, with a narrow distribution concentrated at the lower end. In contrast, the "25-54 years" and "55 years and over" groups show higher wages and broader distributions, reflecting greater variability in earnings as individuals gain experience and qualifications. The older age group generally earns the highest wages, highlighting the strong association between age, experience, and earning potential.

3 Model

Our modeling approach seeks to quantify the relationship between demographic and educational factors and average hourly wages in Canada. To achieve this, we employ a linear regression model to examine how predictors such as education level, gender, and age group influence hourly earnings. The model is implemented using the `stan_lm` function, with a Gaussian distribution to capture the variability in hourly wages.

In this analysis, we focus on predictors that capture key socio-economic and demographic characteristics. Specifically, we include `Education_level`, a categorical variable representing individuals' highest level of education; `Gender`, indicating whether the individual identifies as male or female; and `Age_group`, categorized into "15-24 years," "25-54 years," and "55 years and over." These predictors allow us to explore how differences in education, gender, and age intersect to shape wage outcomes.

The model assumes that the average hourly wage, given these predictors, follows a normal distribution. This Gaussian assumption simplifies parameter estimation and aligns with standard practices in wage modeling. Additionally, we assume moderate priors to avoid overfitting while ensuring interpretability of the coefficients. This balanced approach enables us to identify meaningful relationships between the predictors and wages. Background details and diagnostics are included in Appendix C.

3.1 Alternative model

Initially, the `Education_level` variable was grouped into broader categories, such as "Low," "Medium," and "High" education levels, to simplify the analysis. This approach aimed to reduce the model's complexity while capturing general trends in wage variation. However, this categorization diminished the model's ability to detect differences in wages associated with specific education levels. For instance, combining "Bachelor's degree" and "Above bachelor's degree" into a single category masked the wage attributed to higher education, resulting in a less precise analysis.

An alternative approach was also explored by excluding the `Age_group` variable, based on the hypothesis that its effects might overlap with those of education and gender. However, this exclusion led to a poorer model fit, as it failed to account for wage differences across age groups. Retaining `Age_group` as a distinct predictor improved the model's performance, offering a more accurate depiction of how wages vary across life stages.

3.2 Model set-up

The model predicts the average hourly wage using the following predictor variables:

- Education Level (`Education_level`): A categorical variable representing the highest level of formal education attained by an individual. Levels range from "0–8 years" to "Above bachelor's degree."
- Gender (`Gender`): A binary variable indicating whether an individual identifies as "Male" or "Female."
- Age Group (`Age_group`): A categorical variable representing the age ranges of individuals, categorized as "15–24 years", "25–54 years", and "55 years and over".

The model takes the form:

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{Education level}_i + \beta_2 \cdot \text{Gender}_i \\
&\quad + \beta_3 \cdot \text{Age group}_i + \epsilon_i \\
\epsilon_i &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

Where:

- β_0 is the intercept term.
- $\beta_1, \beta_2, \beta_3$ are the coefficients for each predictor.
- σ^2 is the variance of the error term.

The model is executed in `R` (R Core Team 2023) using the `rstanarm` package (Goodrich et al. 2022). Default priors from `rstanarm` (Goodrich et al. 2022) are used, with the priors set to have a mean of zero and a moderate standard deviation to ensure a reasonable level of regularization.

3.3 Model justification

Existing economic theories and labor market research suggest that education level, age group, and gender notably influence wage outcomes. Higher levels of education are associated with greater specialization and skills, which typically result in higher earnings. Age reflects work experience and career progression, with older individuals often earning more due to accumulated skills and seniority. Gender remains a key determinant, as wage disparities between males and females persist across various labor markets. These predictors collectively capture essential socio-economic dimensions that shape wage structures in Canada.

A Bayesian linear regression model was selected because the dependent variable (average hourly wage) is continuous and approximately normally distributed. This model is well-suited for assessing the contribution of each predictor to wage outcomes while controlling for the effects

of others. For instance, the coefficients of education level, age group, and gender directly indicate their respective impacts on average wages.

Further justification for using this model comes from its alignment with the central limit theorem, as the data aggregates wage observations across individuals. Additionally, the predictors align with established labor market theories, giving a solid theoretical underpinning to our model. The inclusion of priors in the Bayesian framework prevents overfitting, balancing interpretability and predictive accuracy.

A key limitation of this model is that it is entirely trained on the analysis dataset, without splitting into training and testing subsets. Each observation in the dataset represents a unique combination of education level, age group, and gender, making it impossible to partition the data without losing critical combinations. While this approach ensures that all available data contribute to parameter estimation, it restricts the ability to validate predictions on unseen data. To compensate, internal validation methods, such as posterior predictive checks, are employed to evaluate the model's fit and generalizability. Further discussion on this issue is provided in Section 5.

4 Results

4.1 Model Interpretation

4.2 Findings

5 Discussion

5.1 Findings

5.2 Limitations and Future Research Directions

Appendix

A Idealized methodology and survey

B Data Manipulation and Cleaning

During the data cleaning phase, the R packages `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), and `arrow` (Richardson et al. 2024) were utilized. The raw data was imported using `read_csv` from the `tidyverse` (Wickham et al. 2019) package. Subsequent operations filtered the dataset to focus on Canadian data, both full-time and part-time. Additionally, the focus was placed on records that reported the average hourly wage rate, while entries with ambiguous educational levels such as "Some high school" and aggregate categories like "Total, all education levels" were excluded to maintain data clarity and relevance.

The analysis was specifically targeted at the demographic group aged between 25 and 54 years, as this range represents a prime working-age population. Columns important to the analysis were chosen and renamed for better readability and straightforward reference in subsequent analytical procedures. This process included renaming the columns for the year and average hourly wage rate to 'Year' and 'Avg hourly wage rate' respectively.

Education levels were organized in a logical order ranging from "0 - 8 years" of education to "Above bachelor's degree" using the `factor` function to convert them into a categorical variable with a specified level order. This categorization was further enhanced by creating an accompanying numeric variable that mapped these ordered education levels to integers, thus facilitating quantitative analysis.

The final step involved saving the cleaned and structured data. This was done using the `write_csv` function to generate a CSV file for broad compatibility and the `write_parquet` function from the `arrow` (Richardson et al. 2024) package for a more compressed and efficient file format, both of which were stored in the `data/analysis_data` directory.

In all figures and tables, the library `here` (Müller 2020) was used to ensure that the file path should be accessible in all directories.

C Model details

C.1 Posterior predictive check

C.2 Diagnostics

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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://github.com/apache/arrow/.

Statistics Canada. 2020. "Wages by education level." *Open Government Portal*. https://open.canada.ca/data/en/dataset/1f14addd-e4fc-4a07-9982-ad98db07ef86.

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

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