Degrees of Difference: Analyzing the Impact of Education on Earnings in Canada*

Yuanyi (Leo) Liu

April 19, 2024

In this paper, we explore the connection between education and hourly wages in Canada, focusing on data from the year 2000. Our findings indicate a clear trend: higher educational levels correlate with increased average hourly wages for individuals aged 25 to 54. This research highlights the importance of education in determining earning potential and suggests that investment in education could have long-term economic benefits. The study provides evidence for policy implications regarding educational incentives and workforce development strategies in Canada.

Table of contents

*Code and data are available at: https://github.com/leoyliu/Analyzing-the-Impact-of-Education-on-Earnings-in-Canada

1 Introduction

The relationship between education and earnings is a well-established topic of interest within the field of labor economics, providing insights into the broader socio-economic fabric of a country. As Canada's economy continues to diversify and specialize in various sectors, the value of education in this landscape remains a critical question for policymakers and the public. While previous studies have explored this link, there is an ongoing need to update and deepen our understanding of how this dynamic plays out in the modern economy.

This paper narrows the focus to the Canadian labor market, where the interplay between educational achievement and wages within the core working age group, those aged 25-54, offers a mirror to the efficacy and value of educational advancements in the workplace. This demographic is selected for its relevance to the labor force and policy implications, excluding younger individuals still likely in the education system and older individuals, who may present outlier educational experiences that do not align with the central workforce.

This investigation seeks to address a gap in the current literature by providing a targeted analysis of the 25-54 age group in Canada since the year 2000, a period marked by rapid technological change and economic evolution. By applying a linear model to wage data categorized by education level, this study uncovers a positive correlation between educational attainment and hourly wages, suggesting that higher education can be linked to improved wage outcomes in this key demographic.

The structure of the paper is organized as follows: Following Section 1, Section 2 presents the data, detailing the data sources, analytical techniques, and the rationale behind the chosen methods. Section 3 then covers the specifics of the linear model analysis, laying out the statistical underpinnings that support our investigation. After that, Section 4 discusses the results, elaborating on the observed trends and patterns in wage rate data. Section 5 interprets these findings in light of the current economic and educational context in Canada, exploring potential factors influencing these trends, drawing connections to broader socio-economic issues, and providing suggestions for future research in this area.

1.1 Estimand

In this study, our primary focus is to estimate the effect of education on hourly wages in Canada. Our estimand is the incremental average hourly wage rate for individuals within the specified education levels, compared to the next lower education level. By quantifying this effect, we aim to capture the economic value of educational attainment and how it translates into wage premiums.

2 Data

This section aims to offer an insightful understanding of the dataset utilized in our analysis, which serves as the foundation for our examination of the relationship between education and average hourly wage rates in Canada, specifically within the 25-54 age demographic.

2.1 Source and Methodology

This study uses a dataset sourced from the Open Government Portal of Canada (Statistics Canada 2020), specifically designed to track the correlation between educational attainment and hourly wages across various demographic segments of the Canadian workforce. The dataset covers the period from the year 2000 onwards, providing a longitudinal view of wage trends in relation to educational background.

Alternative datasets, such as those from Statistics Canada's Labour Force Survey, were considered but were not selected due to their less detailed categorization of education levels and their broader focus on employment without specific wage breakdowns by education.

The data was processed and cleaned using R(R Core Team 2020), a powerful statistical programming language. The selection of this dataset was motivated by its direct relevance to the research question and its regular updates, ensuring that the data remains relevant for observing current trends. Initial data processing involved filtering the dataset to focus solely on entries categorized under 'Canada' in geography, eliminating data irrelevant to the national focus of this research. This step was crucial to maintaining the clarity and relevance of the analysis. For key operations, please refer to the Appendix A.

2.2 Variables

To better understand data, key variables extracted for this study include the year of data collection, education level, age group, and the average hourly wage rate. Notably, we have streamlined the 'Education level' from a broad range of categories into a numeric variable that aligns with ascending educational attainment, facilitating a quantitative analysis of its impact on wages. The age group has been limited to the 25-54 years range to focus on the

most economically active segment of the population, avoiding potential outliers from younger individuals with less work experience and older individuals whose advanced education, such as doctoral degrees, could distort the analysis. Wage rates are presented in Canadian dollars per hour, reflecting pre-tax earnings, inclusive of tips, commissions, and bonuses.

Table 1: First Ten Rows of Hourly Wage Across 2000-2019 After Data Processing

Year	Education Level	Age	Hourly Wage
2000	0 - 8 years	25-54 years	13.1
2000	High school graduate	25-54 years	15.8
2000	Post-secondary certificate or diploma	25-54 years	18.0
2000	Trade certificate or diploma	25-54 years	17.4
2000	Community college, CEGEP	25-54 years	18.1
2000	University certificate below bachelors degree	25-54 years	20.2
2000	University degree	25-54 years	23.4
2000	Bachelor's degree	25-54 years	22.5
2000	Above bachelor's degree	25-54 years	25.6
2001	0 - 8 years	25-54 years	13.3

Table 1, created with `kableExtra` (Zhu 2021), outlines the first ten records of our dataset, featuring average hourly wages for Canadians aged 25-54 across varying levels of education from the years 2000 to 2019. This snapshot reveals the variable structure of our analysis, which includes the `Year` of wage data, the defined `Education level` ranging from "0 - 8 years" to "Above bachelor's degree," and the consistent `Age group` focus. It provides an early indication of the ascending trend in wages with higher educational attainment, setting the stage for our deeper investigation into the economic value of education within Canada's workforce.

To transition from a tabular overview to a more graphical representation, we will visualize the data, which will allow us to spot patterns and trends that are not immediately apparent in numerical form.

Figure 1 built by `ggplot2` (Wickham 2016) illustrates the distribution of average hourly wage rates across different educational levels from 2000 to 2019 for Canadian workers aged 25-54. Each dot represents the wage rate for a specific education level in a given year, with color coding to distinguish between educational categories, ranging from "0 – 8 years" to "University degree." The spread and upward trend of the points suggest that individuals with higher educational qualifications tend to have higher average hourly wages, and this trend persists over the 20-year period.

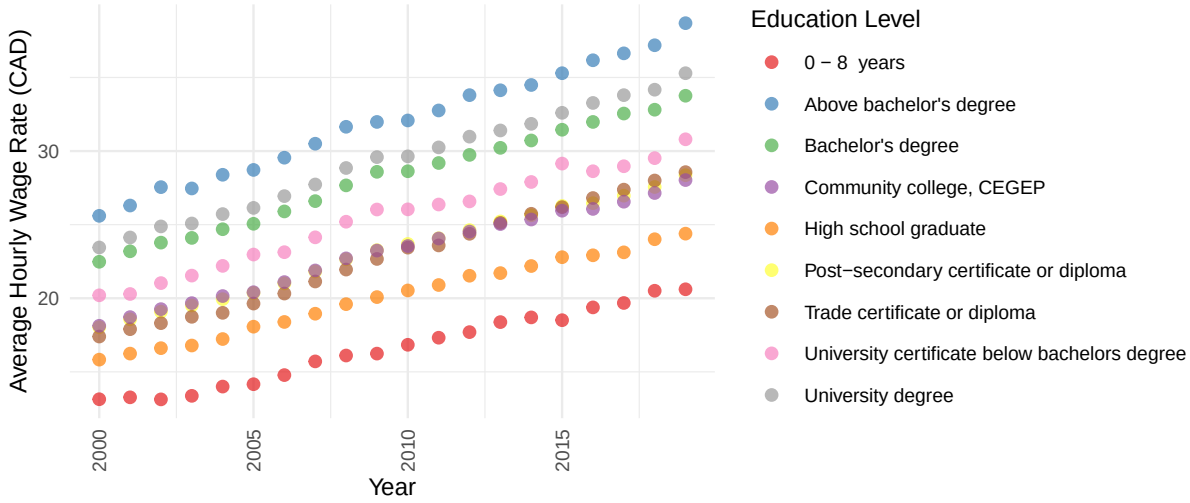


Figure 1: Average Hourly Wages by Educational Attainment, 2000-2019

2.3 Measurements

The measurement of variables within this dataset was handled to ensure accuracy and relevance. The 'Average Hourly Wage Rate' is calculated based on actual wages reported by employers, providing an objective measure of income. This data is then vetted and standardized by the Government of Ontario to ensure it reflects accurate and fair representations of wages across different demographics.

For educational attainment, data entry occurs as individuals enter the workforce or update their qualifications, with categorizations reflecting the highest level of formal education completed. This data is collected through surveys and employment records, often as part of broader demographic data collection efforts by governmental agencies.

Correlations between the variables will be examined at Section 3, providing insights into how education levels are associated with hourly wages. By understanding these relationships, we can better infer the potential impact of education on earnings within the Canadian economy.

3 Model

The objective of our modeling approach is to quantify the relationship between educational attainment and hourly wages. Our analysis employs a Bayesian framework to assess how changes in the level of education correlate with variations in wage rates among Canadian workers aged 25-54.

In our analysis, we utilized a simple linear regression model, a technique well-suited for examining the relationship between a continuous dependent variable and one or more independent

variables. Given that our dependent variable, `average hourly wage rate`, is continuous, a normal or Gaussian linear regression model is appropriate. This model assumes that the distribution of the wage rate, given the level of education, follows a normal distribution. This assumption allows for a straightforward interpretation of the parameters and is commonly used in practice. The use of a normal distribution for the error term is a standard choice, as it facilitates the estimation of the model parameters using maximum likelihood estimation. The simplicity and interpretability of this model, coupled with its appropriateness for the data at hand, make it a good fit for our study's objectives.

Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i represent the average hourly wage rate for the i^{th} individual. The predictor variable, x_i, corresponds to the numeric value assigned to each education level. The model can be described by the following equations:

$$\begin{aligned} y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \beta_0 + \beta_1 x_i \\ \beta_0 &\sim \text{Normal}(0, 2.5) \\ \beta_1 &\sim \text{Normal}(0, 2.5) \\ \sigma &\sim \text{Exponential}(1) \end{aligned}$$

The model is executed in R (R Core Team 2020) using the `rstanarm` package (Goodrich et al. 2024). Default priors from `rstanarm` (Goodrich et al. 2024) are used, set with a mean of zero and a conservative standard deviation.

3.2 Model justification

Given the economic theory and previous findings, we anticipate that higher educational qualifications would correspond to higher hourly wages. This is encapsulated in the assumption that β_i should have a positive effect on y_i. Through Bayesian analysis, we aim to capture the degree to which educational attainment can predict wage rates, thus providing evidence to support or refine this hypothesis.

We selected a Gaussian linear regression model to examine the relationship between education level and average hourly wages because our dependent variable, the wage rate, is a continuous data point that typically follows a normal distribution. Linear regression is a straightforward method for assessing the impact of one or more independent variables on a dependent variable. In this case, it helps us understand how increasing education levels, which we have numerically encoded, relate to wage rates.

Table 2: Summary of the Linear Regression Model

	Linear Model
(Intercept)	16.143
Education_numeric	1.672
Num.Obs.	180
R2	0.610
R2 Adj.	0.607
Log.Lik.	−478.097
ELPD	−480.2
ELPD s.e.	7.1
LOOIC	960.5
LOOIC s.e.	14.2
WAIC	960.4
RMSE	3.43

Further justification for this model comes from its foundation in the central limit theorem, which suggests that with a large sample size, our wage rate data should be normally distributed. Moreover, the expected positive correlation between education and wages aligns with existing economic studies, reinforcing our theoretical framework. The simplicity of the model aids in avoiding overfitting, ensuring that the results are generalizable and relevant to a broader population. The regression model's assumptions—linearity, independence, and normality of residuals—are all conditions that our dataset reasonably meets, making a Gaussian linear regression not only a convenient choice but also statistically appropriate for our analysis.

4 Results

Section 4 explores the relationship between education and average hourly wages in Canada. Utilizing a dataset that captures 20 years of wage data across varying educational levels, we apply a linear modeling approach to uncover patterns and draw conclusions about the economic value of educational advancement. Below, we present the results of our model and its implications.

The results obtained from our linear model, which predicts the average hourly wage rate based on the numeric representation of education levels, show a clear and positive relationship between educational attainment and wage rate. Table 2 indicates that the intercept is estimated at 16.143, suggesting that when the education level is at the baseline (0 - not accounted for in the numeric scale), the average hourly wage rate would be approximately 16.14. Importantly, the coefficient for Education_numeric is positive (1.672), supporting the hypothesis

that higher educational levels correlate with higher wage rates. The model used 180 observations for the analysis, with an R-squared value of 0.610, which means that approximately 61% of the variance in the average hourly wage rate is explained by the model. The adjusted R-squared, which accounts for the number of predictors in the model, is very close to the R-squared, indicating a good fit without overfitting. Notably, the RMSE (Root Mean Square Error) of 3.43 reflects the standard deviation of the prediction errors, which measures how spread out these errors are around the true regression line.

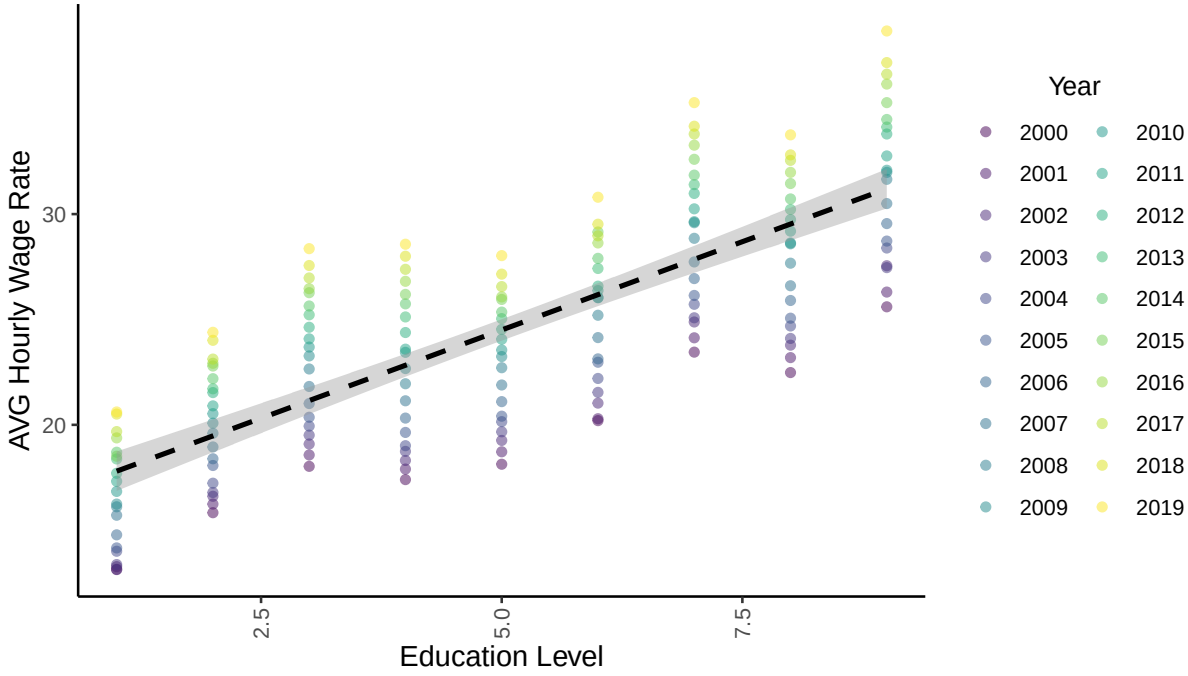


Figure 2: Relationship between Hourly Wage Rate and Education Level

Figure 2 illustrates the fitted linear regression line across the spread of average hourly wage rates for different education levels over the years 2000 to 2019. Each point in the scatter plot corresponds to the average hourly wage rate for a given education level, differentiated by color to represent each year. The dashed line across the plot signifies the best-fit linear regression line, which inclines upward as the education level increases. The diversity of colors along the regression line also reveals the variation in wage rates across different years, although the increasing trend is consistent throughout.

In conclusion, the results presented align with the initial hypothesis, showing that as education level increases, there is a corresponding increase in the average hourly wage rate. This validates the underlying assumption that more advanced education tends to lead to better-paying job opportunities. The analysis effectively captures the essence of the data, allowing for a meaningful interpretation that leads us into the discussion. In the next Section 5, we will discuss the implications of these findings, examining their relevance to current labor market

theories and the potential for informing education and economic policy decisions.

5 Discussion

In this paper, we investigated the relationship between educational attainment and average hourly wage rates from 2000 to 2019. By deploying a linear Bayesian model, we have been able to highlight the correlation between higher educational attainment and increased wage rates, aligning with our original hypothesis.

5.1 Findings

Our analysis reveals a clear, positive correlation between education level and average hourly wage rates across the data spanning from 2000 to 2019. The consistent upward trend in wage rates with higher education levels across the years underscores the value of educational attainment in the labor market. Despite some fluctuations, the general pattern suggests that investments in education are likely to be economically beneficial for individuals in the age group of 25-54 years.

5.2 Insights on Education and Wages

The results of this study offer clear evidence that educational attainment is closely tied to wage earnings, highlighting the premium that the labor market places on higher qualifications. This finding is consistent with current labor market theories that emphasize the importance of human capital—suggesting that investments in education are indeed reflected in higher wages. Specifically, our data supports the view that a well-educated workforce is critical for economic growth and competitiveness in today's globalized world, which increasingly values knowledge and specialized skills.

From a policy perspective, these insights could serve as a catalyst for reform, underscoring the need for investments in education. Policymakers might use this information to support sectors that are likely to offer greater economic returns for educational investments. This could involve bolstering STEM (Science, Technology, Engineering, and Mathematics) education, for example, or providing incentives for pursuing in-demand fields such as healthcare or information technology.

In conclusion, this study's findings highlight the significance of education as a determinant of economic success and reinforce the importance of informed educational and economic policymaking. For policymakers, the imperative is clear: to design and implement educational strategies that not only increase access to education but also ensure that the education provided is relevant and responsive to the current and future needs of the labor market.

5.3 Limitations and Future Research Directions

The primary limitation of this study is its relatively small sample size, with only 180 observations. This constraint may impact the generalization of the findings. Additionally, the focus on the age group of 25-54 years might not encapsulate the full spectrum of the workforce, especially considering the varied career stages and life circumstances of younger and older workers. Such a limited scope might overlook the effects of education on wages across different life stages.

Future research should aim to disaggregate the age groups to understand the differential impact of education on wages. Distinguishing between early-career individuals, mid-career professionals, and late-career individuals could yield more insights into the education-wage dynamic at different life stages. Additionally, expanding the sample size and including more varied age groups would enhance the representativeness and depth of the analysis.

Appendix

A 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.

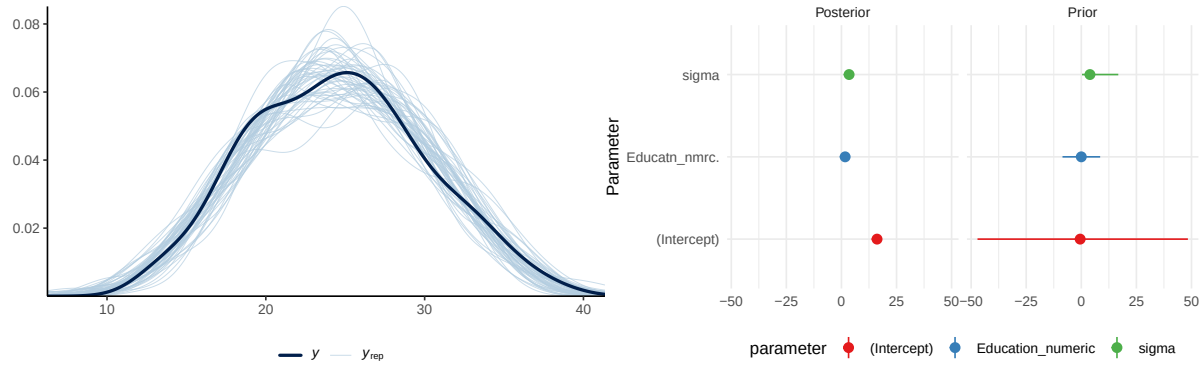
B Model details

B.1 Posterior predictive check

In Figure 3a, we implement a posterior predictive check, which displays the overlap between the observed data (denoted by y) and the replicated data generated from the model (denoted by y_{rep}). The density of the replicated data largely coincides with the observed data, suggesting that our model does a reasonable job of capturing the underlying distribution of the observed average hourly wage rates. This alignment is crucial as it indicates that the model's predictions

are consistent with the real-world data we aim to understand and that the model can be a reliable tool for further inference.

In Figure 3b, we compare the posterior distributions of our parameters with the priors. This comparison reveals how the observed data have updated our beliefs about the parameters. We see that the posterior distributions for the intercept and the education_numeric coefficient are more concentrated and shifted from the priors, indicating that the data provided substantial information to update our initial assumptions. The sigma parameter's posterior, which represents the standard deviation of the residuals, is also refined from its prior, showing that the data have informed us about the variability in wages that is not explained by education level alone.

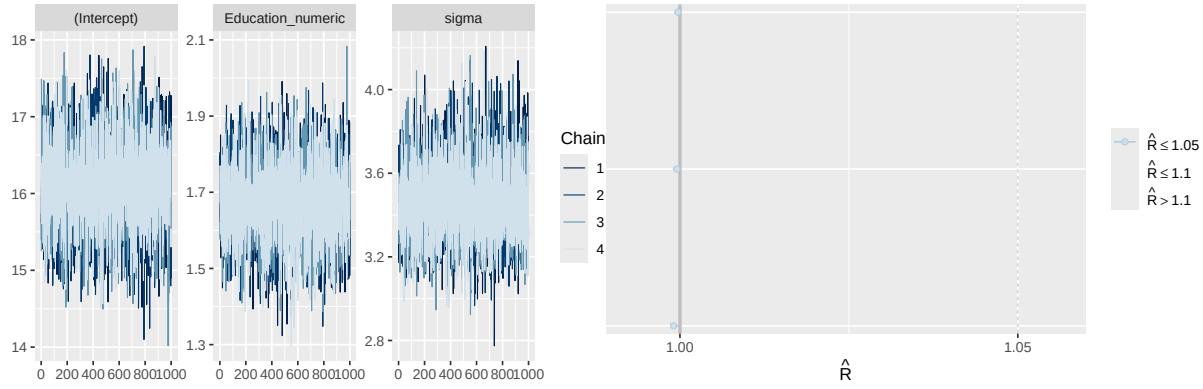


(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 3: Examining how the model fits, and is affected by, the data

B.2 Diagnostics



(a) Trace plot

(b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

In Figure 4a, the trace plot demonstrates the sampling paths of the Markov chains for each parameter in the model. The trace for the intercept, education_numeric coefficient, and sigma parameter all display a 'hairy caterpillar' shape, indicating good mixing and convergence across multiple chains. This suggests that the model's parameter estimates are stable and reliable, as evidenced by the overlapping and intertwined paths, which is indicative of a well-fitting model.

Figure 4b, the Rhat plot, presents the R-hat statistics for each parameter, which is a measure of convergence. We see that all values at or very close to 1.00. This is a desirable outcome, implying that further sampling would likely yield little additional information and that the posterior estimates can be considered reliable for inference.

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "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. 2020. *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. 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, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.