Determinants of Wage Variation in Canada*

The Influence of Education, Gender, and Age on Canadian Earnings

Yuanyi (Leo) Liu

November 26, 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: Determinants of Wage Variation 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.

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.

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.

2 Data

This section aims to offer a 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 Overview

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 Appendix C.

2.2 Measurement

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 in 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.

2.3 Outcome variables

2.3.1 Hourly Wages

2.4 Predictor variables

2.4.1 Education

2.4.2 Gender

2.4.3 Age

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 D.

3.1 Alternative model

3.2 Model set-up

The model predicts hourly wages using the following predictor variables:

-

The model takes the form:

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

The model is executed in `R` (R Core Team 2020) using the `rstanarm` package (**rstanarm?**). Default priors from `rstanarm` (**rstanarm?**) 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

4 Results

5 Discussion

5.1 Findings

5.2 Insights on Education and Wages

5.3 Limitations and Future Research Directions

Appendix

A Census methodology overview and evaluation

B Idealized methodology and survey

C 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.

D Model details

D.1 Posterior predictive check

D.2 Diagnostics

References

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, 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.