Income is a Strong Predictor for Higher Education Enrolment

Matthew Wankiewcz, Xiaoyan Yang, Alen Mitrovski, Harry Hwang

19/10/2020

Abstract

The need for higher education is an idea that is constantly growing around the world, especially in Canada. In this paper, we run a logistic regression model to determine which factors seem to impact someone's chances of only having a high school diploma using GSS survey data from 2017. From our model, we find that the higher a person's income is, the more likely they are to have completed another level of education other than high school. These results are significant as it shows that a high school education may not lead to high incomes in Canada.

Introduction

The role of education is to drive many aspects of development. School education provides everyone with the opportunity to improve their literary ability, obtain scholastic and professional skills, and to explore the unknown world. In the adult world, education is the main approach to divide people into different occupations, financial statuses, and even different social classes. Our curiosity leads us to ask: what are the main factors connected to people's education level in Canada? The Canadian General Social Survey (GSS) on Family provides us with an opportunity to study these relationships. The General Social Surveys are designed and performed by statistics Canada on an annual basis (GSS). We utilize GSS on Family in 2017 survey data as our research data.

The data of GSS on Family in 2017 was released in February 2019 the link can be found here. The main objective of GSS on Family is to monitor changes in Canadian families. The target population for the 2017 GSS is all non-institutionalized individuals, aged 15 years or older, living in the 10 provinces of Canada. The survey includes 81 variables. These variables contain information on conjugal and parental history, family origins, children's home leaving, fertility intentions, and other socioeconomic characteristics (GSS). To achieve our objective, we selected sex, married/nonmarried, respondent's annual income, born in Canada or outside Canada as the main explanatory variables to investigate their relevance with education level. We divided the highest education level completed into two categories: high school diploma and education above high school. Education above high school includes college education, trade certificates, bachelor's degree, master or above. Since GSS on Family was not designed for an education-based study, the obtained data limited our research's breadth and depth. We obtained 19 423 data responses and applied multilevel regression methodology in our study. The selected variables are strongly correlated with education levels and the regression p-values support the significance of the calibration results. Due to time and data limitations, we were unable to perform poststratification adjustments and goodness fit tests for our model. More specific questions in the survey are required to enhance this study. These deficiencies will be our improvement for future analysis.

Data

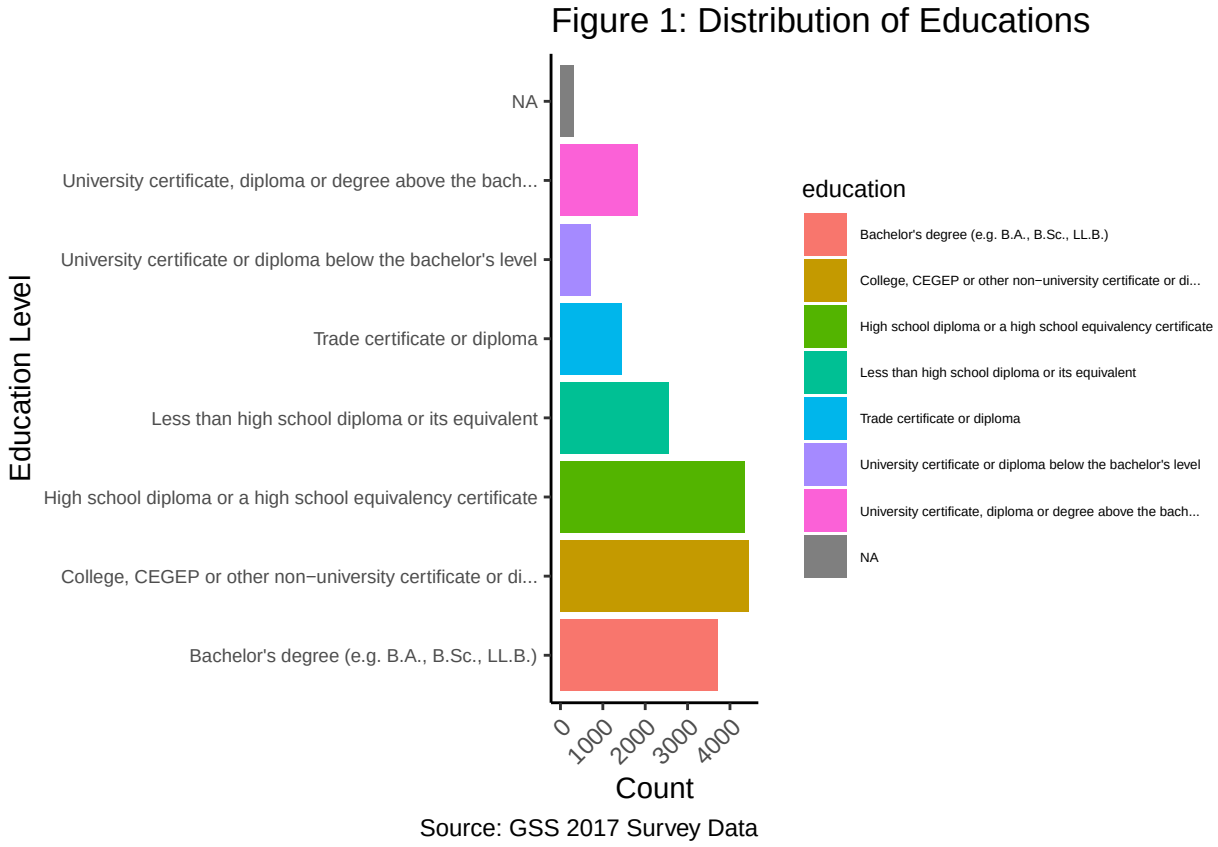
The dataset we used was from the 2017 GSS survey conducted by the Canadian Government [1]. According to the Government's website, the objectives of the survey were to gather data on social trends and to provide information on the most important social issues.

A target population is defined as the set of all units covered by the main objective of the study (Wu, 2020). In this study, the target population was the general Canadian population.

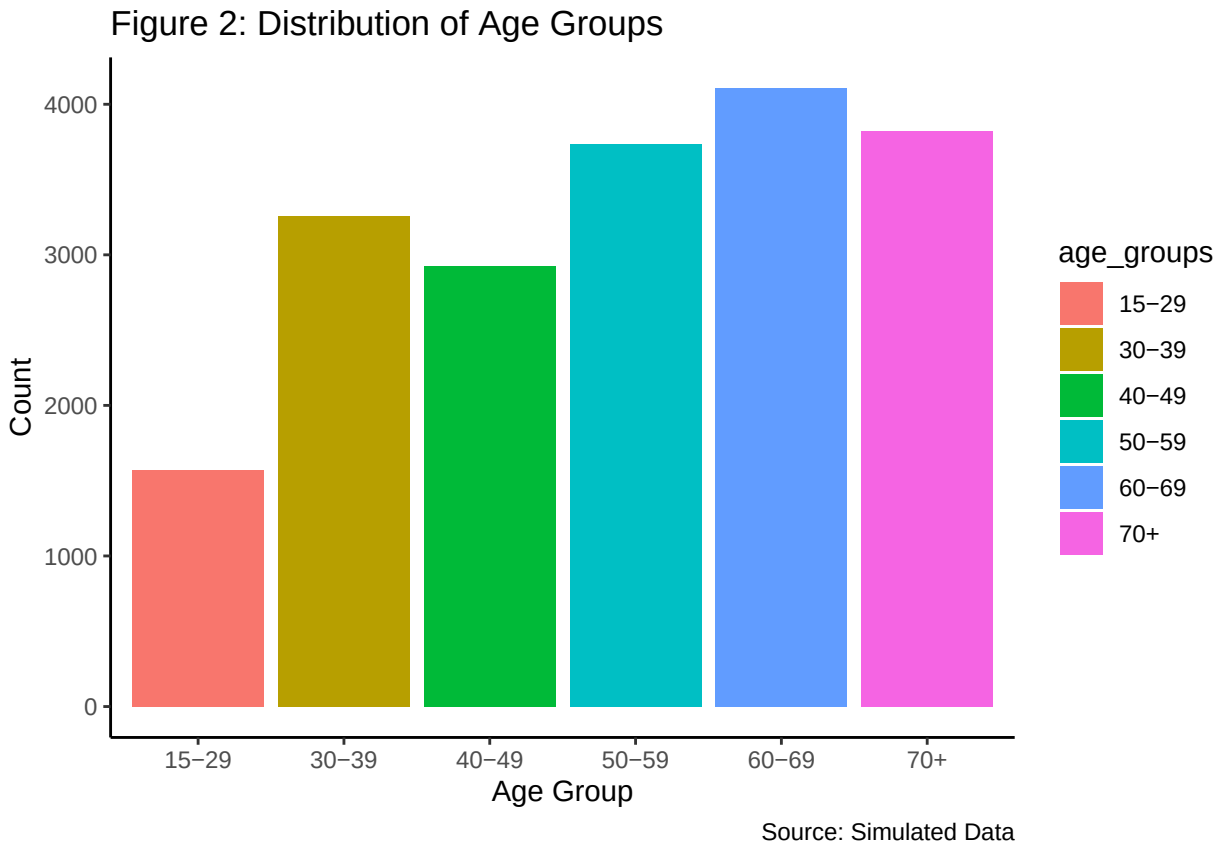
A sampling frame is defined as the list of units in the survey population (Australian Bureau). The sampling frame for this study was individuals aged 15 or older in the various Canadian provinces and territories.

A sampled population, also known as a study population, is defined as the population represented by the survey sample (Wu, 2020). The sampled population in this GSS study is represented by the 19 423 individuals that completed this survey

Respondents were found through the Census because it is conducted on people over the age of 15, it fits right in with their target population. Their initial sample had 43,000 people and brought down the number of invitations sent to 34,000 households while 20,000 responses were expected. The government reported a response rate of 52.4% and with our records, they collected 20602 responses. As for non-response, the researchers changed the weighting on responses (making some responses account for more than others) in order to account for non-response to avoid bias.



[1]To obtain the data, we used the GSS cleaning script from Rohan Alexander and Samantha-Jo Caetano. This file is included in the GitHub repo which supports this analysis and can be found here. GSS files can be accessed by going to www.chass.utoronto.ca, then clicking data centre, signing in, and finding the GSS survey. Then click the desired year (we chose 2017), then click download csv in STATA form, select all variables and download the csv. Then using the gss_cleaning-1.R file in inputs along with gss_dict and gss_labels, run the cleaning script to get your csv.



Figures 1 and 2 show the distribution of education level, and age.

Figure 3: Visible Minority Summary

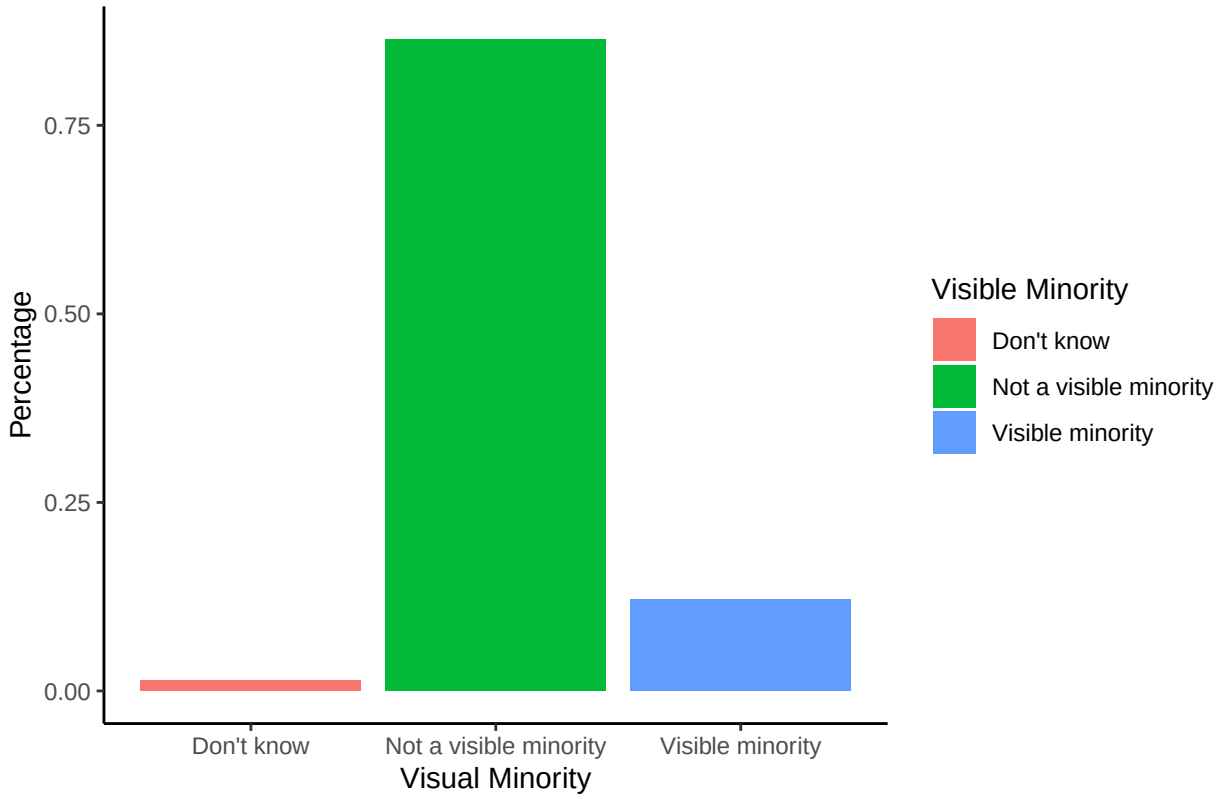


Figure 4: Born in Canada Summary

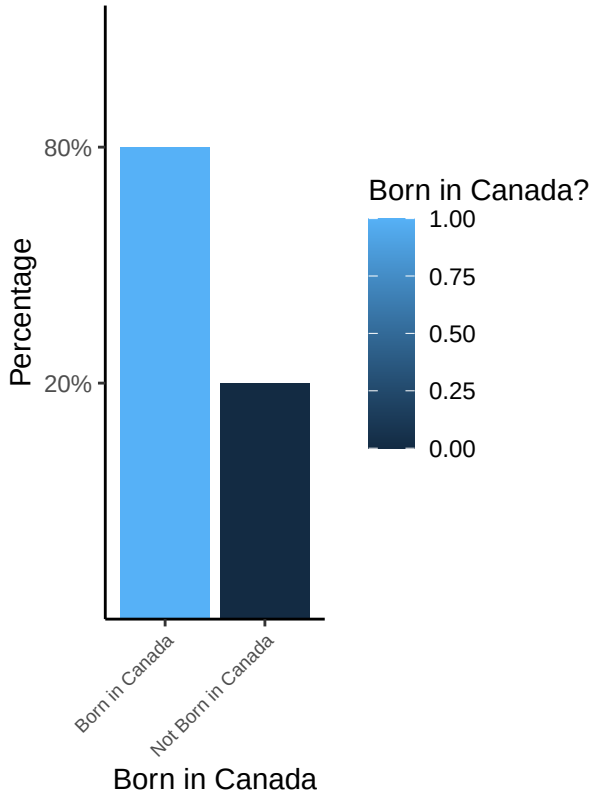


Figure 5: Gender Summary

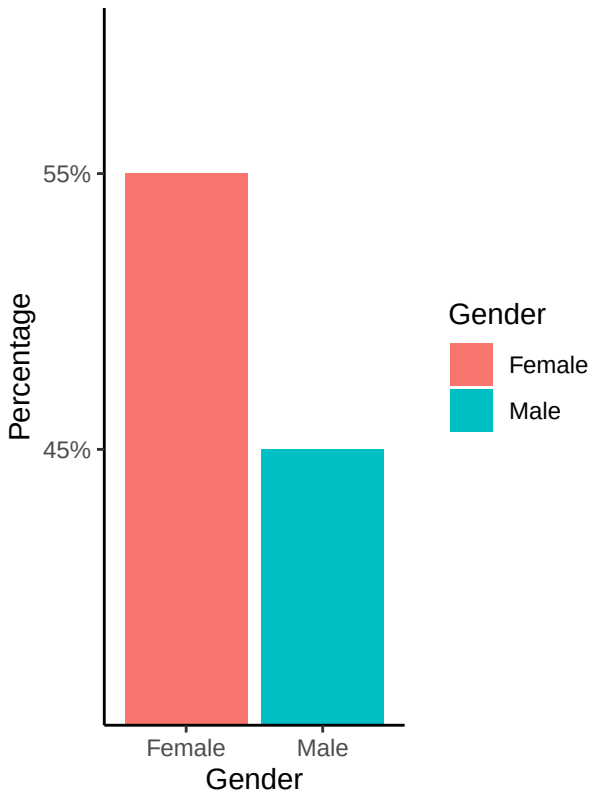


Figure 6: Population by Provinces

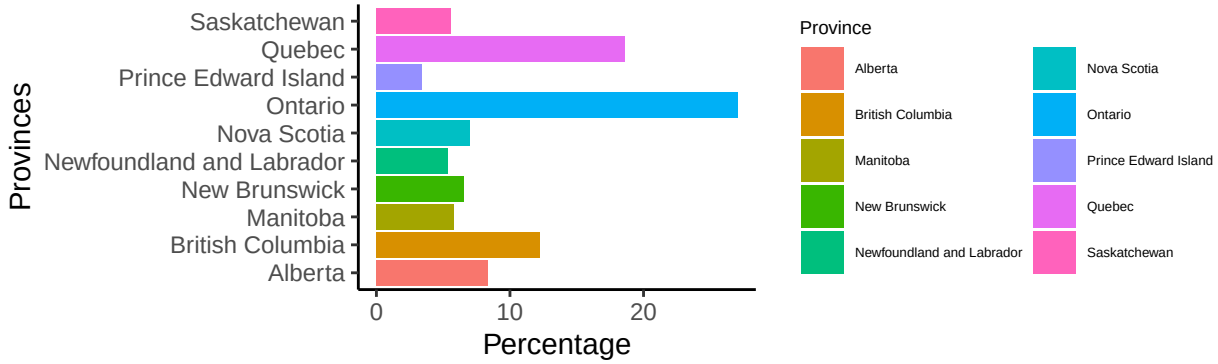
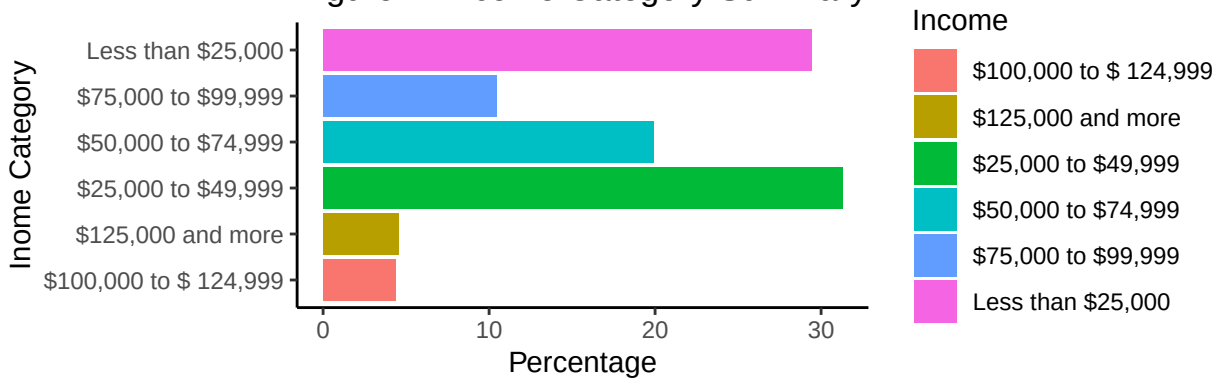


Figure 7: Income Category Summary



Figures 3-7 show various aspects of the demographics of the survey respondents. This includes the proportions of racial minorities, proportions of those born in Canada, the gender proportions, and provincial and income level proportions.

From our plots, we can see some strengths and weaknesses of the survey. We can see that the distribution of age is fairly even which allows us to get a good idea of the opinions of all age ranges. Another strength is that the distribution of education levels seems to be accurate, we see that fewer people tend to get Bachelor's degrees compared to high school diplomas, which bodes well for our intended model, we need a fairly even distribution to get a good model.

As for our weaknesses, we see that many respondents had incomes under 50 000 dollars, which could represent people that are retired or younger people who are in school and don't work as much. Also, some questions are subjective in this survey, such as self-rating your health and mental health or your feelings on life. This is tough to interpret because one person's idea of good health could be someone's definition of poor health.

For our model, we used the variables: number of children, sex, respondent income, whether the respondent is married not, whether the respondent was born in Canada and lastly we made a new variable which is binary and gives a 1 if the respondent has only completed high school or equivalent and 0 otherwise.

Model

We plan to run a logistic regression model in the form:

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_1 + \beta_2 * x_{children} + \beta_3 * x_{male} + \beta_4 * x_{income>125000} + \beta_5 * x_{income_{25000-49999}}$$

$$+\beta_6 * x_{income_{50000-74999}} + \beta_7 * x_{income_{75000-99999}} + \beta_8 * x_{income_{<25000}} + \beta_9 * x_{marraige} + \beta_{10} * x_{canada}$$

where each β represents a coefficient determined by our model and is multiplied by each variable we planned to include. For some variables such as income, we see x values for each level. This happens because our model will expect a 1 or 0 for each of the entries, so if someone has an income under 25 000 dollars, that β will be 1 and the others for income will be 0. Also, the β for number of children is just multiplied by how many children the person has, it does not use binary values like the other coefficients. $\hat{p}$ represent the probability a respondent has completed high school and *log* represents the natural logarithm.

The value we want to predict is whether or not someone has had a higher education, more specifically, we restrict it to a variable which gives a 1 for if a person has only completed high school and a 0 if they've completed some other level of education.

This model was run using the glm function. The glm function was run with the family being specified as binomial() because our response variable "only_highschool" is a binary variable with responses of "yes" or "no".

The ouput of our model gives us a probability of whether someone has only completed high school. We find this by assigning each β value a 1 or 0 (or number for number of children) and then solve it using the equation:

$$\hat{p} = \frac{e^{sum}}{1 + e^{sum}}$$

.

Where e is the exponential function and sum is the sum of our model's equation.

This type of model is best suited for the analysis we want to conduct because it allows us to create a regression line for a binary variable. Instead of having a straight regression line going through some of the points, logistic regression attempts to make a more curved line to capture as many points as possible. We feel that this model is more suited than other models like linear regression because of how versatile it is in being able to change its output depending on a wide variety of inputs. One part where this model falls short is that we get a probability from this model and have to interpret it, while linear regression gets an exact value from the equation.

Our model converges to a maximum probability of about 93% of only having a high school diploma. This is done by inputting the maximum values for our regression (10 kids, Man, income under 25 000, born in Canada and unmarried), this is extreme but there is a chance it is possible. Our lower bound for probabilities is about 3.6%. Again this is using the extreme case (Woman, no kids, income over 125 000, not born in Canada and married).

Results

Figure 8: Basic Output of Logistic Regression Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.8679928	0.1131702	-25.3423052	0.0000000	-3.0935789	-2.6496829
total_children	0.2446155	0.0112286	21.7849773	0.0000000	0.2226615	0.2666797
sexMale	0.4412143	0.0332330	13.2764067	0.0000000	0.3761447	0.5064198
income_respondent$125,000 and more	-0.1108059	0.1418003	-0.7814221	0.4345543	-0.3891234	0.1672631
income_respondent$25,000 to $49,999	1.4933085	0.1040008	14.3586298	0.0000000	1.2933403	1.7013504
income_respondent$50,000 to $74,999	0.7576833	0.1069993	7.0811982	0.0000000	0.5515583	0.9713208
income_respondent$75,000 to $99,999	0.2466038	0.1166471	2.1141002	0.0345067	0.0206266	0.4782202

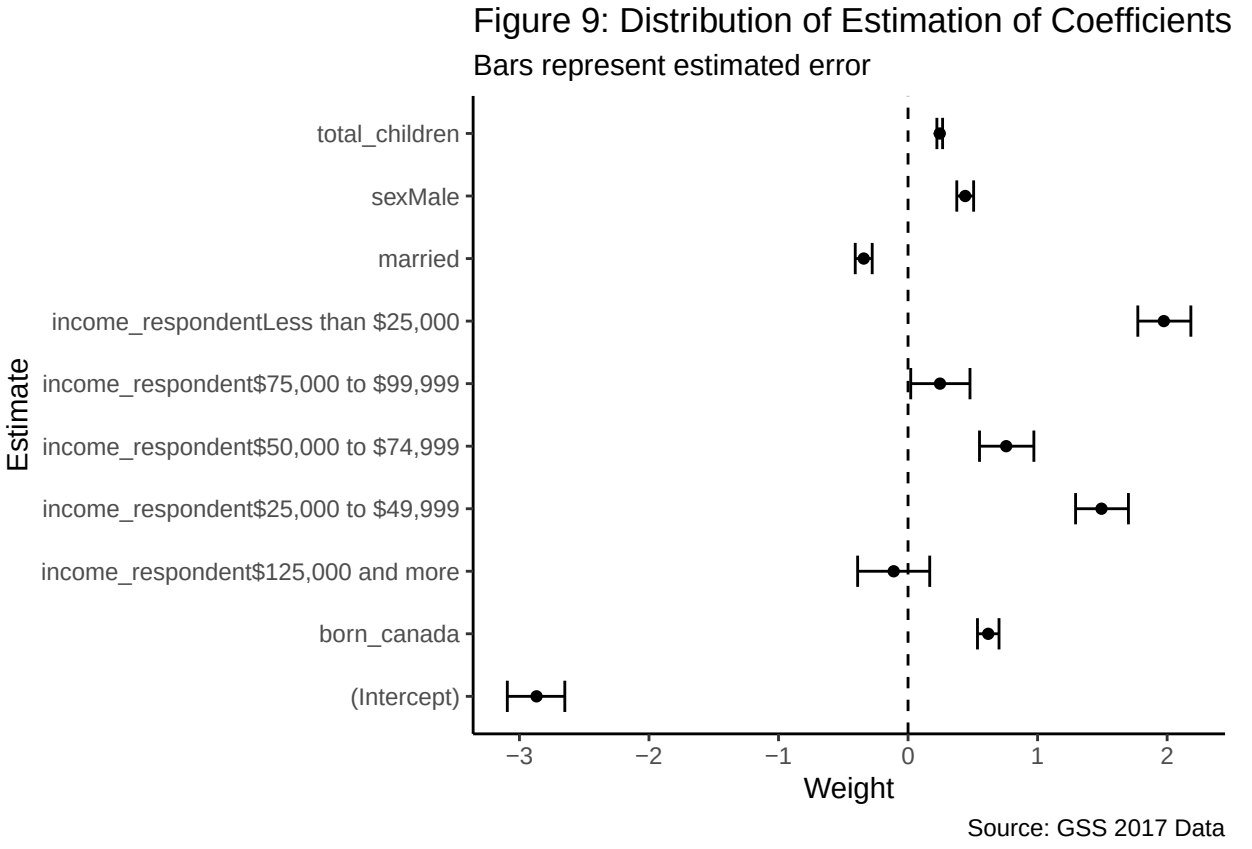
term	estimate	std.error	statistic	p.value	conf.low	conf.high
income_respondentLess than $25,000	1.9745764	0.1044005	18.9134851	0.0000000	1.7738123	2.1833854
born_canada	0.6187816	0.0424360	14.5815180	0.0000000	0.5359124	0.7022717
married	-0.3421994	0.0333583	-10.2583009	0.0000000	-0.4076299	-0.2768634

Using the glm() function, we find that our logit function is:

$$log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.88 + .24 * x_{children} + .44 * x_{male} + -.09 * x_{income>125000} + 1.49 * x_{income_{25000-49999}}$$

$$+.77 * x_{income_{50000-74999}} + .25 * x_{income_{75000-99999}} + 1.97 * x_{income_{<25000}} + .63 * x_{canada} - .34 * x_{marraige}$$

The output of this model is not like other regressions where we plug in numbers and receive our value right away, we have to do a little more work for this one. Every term in this equation will either be multiplied by 1 or 0 if the person has that characteristic or not. You may notice that some categories are missing, such as the female gender, income from 100 000-124 999 dollars and not born in Canada. These are missing because the function takes the other values as default ideal values, so for every difference in the category, the probability of having only a high school diploma is different relative to the baseline values.



This graph shows the distribution of coefficients for each variable covered, with the standard error extended to the left and right. There is a bar in the middle of the graph (at x = 0), to help visualize whether the variable has increases the probability of a person only completing high school.

Figure 10: Logistic Regression Outputs with Confidence Intervals Present

Characteristic	**log(OR)**	**95% CI**	**p-value**
(Intercept)	-2.9	-3.1, -2.6	<0.001
total_children	0.24	0.22, 0.27	<0.001
sex			
Female			
Male	0.44	0.38, 0.51	<0.001
income_respondent			
$100,000 to $ 124,999			
$125,000 and more	-0.11	-0.39, 0.17	0.4
$25,000 to $49,999	1.5	1.3, 1.7	<0.001
$50,000 to $74,999	0.76	0.55, 1.0	<0.001
$75,000 to $99,999	0.25	0.02, 0.48	0.035
Less than $25,000	2.0	1.8, 2.2	<0.001
born_canada	0.62	0.54, 0.70	<0.001
married	-0.34	-0.41, -0.28	<0.001

Figure 10 is a cleaner way of looking at our coefficients, we can find odd ratios of each category impacting whether or not someone has only completed high school. The way we interpret these odds ratios is through the decimal points of the ratios. If one outcome is n decimal places higher, the odds of it impacting the model's decision are n times more likely. For example, we see that the odds ratio for Males is 0.44, which means that males are more likely to have only completed high school. The same idea applies to negative values, it means they are less likely to impact.

For categorical variables like income, the value of the odds ratio depends on whether the value is greater than or less than 1. For values less than one, such as income about 125 000 dollars, it means that a person is less likely to only have a high school education compared to someone with the baseline income of 100 000 to 124 999 dollars. The same idea goes for values above 1, this time it means that the person is more likely to have only completed high school.

We can also look at the p-values, to see which results are the most statistically significant. P-values are used with null hypotheses to determine if an experiment is significant. Since our p-values are under the traditional benchmark of 0.05, we can reject our null hypothesis and conclude that these results are in-fact, significant.

Discussion

Analysis

Now that we've created our model and plotted it, we can interpret what it means and how it impacts the small and large worlds. Firstly, we should interpret our coefficients and how they determine a person's education level. When looking at our p-values for the coefficients, we find that almost all coefficients are very strong evidence to reject our null hypothesis as discussed in Figure 10. We see that having an income level of over 125 000 dollars still gives evidence to reject our null hypothesis (that it has no impact on education level) but it's very close to not being enough evidence. What this means is that although it does decrease your chances of only having a high school diploma, there may have been instances where a person reaches that income level with a high school diploma. This makes sense because in the past people did not need university degrees to make a successful living, because most people didn't go to university. According to an article in the UK Guardian called "Student Experience - Then and Now", "In the early 1960s, only 4% of school leavers went to university, rising to around 14% by the end of the 1970s. Nowadays, more than 40% of young people start undergraduate degrees" (The Student Experience - Then and Now, 2016). This is a significant increase in levels and can help explain the higher than normal p-value for our model's standards.

When examining our data in Figure 8, we see that there is a correlation between having only a high school education and lower income levels. This can be explained through the accessibility to resources, knowledge reservoirs and networking opportunities that are accompanied by post-secondary education. The job market typically values higher education when selecting candidates for positions. According to a 2015 report by Statistics Canada, both men and women with bachelor's degrees as their highest level of education completed earned more income than their counterparts with high school diplomas. Women that completed a bachelor's degree earned roughly 25 000 dollars more than women with only a high school diploma. Likewise, men with bachelor's degrees earned around 26 000 dollars more than those with only a high school diploma (Does Education Pay, 2017).

As seen in Figure 8, men are more likely to have only completed high school as their highest level of education. According to Statistics Canada, education has evolved significantly since 1990, specifically for women in Canada (Women in Education, 2009). The proportion of women that have completed post-secondary education has nearly doubled and the percentage of women who drop out of high school has improved from 26% to 9% over this period (Women in Education, 2009). In contrast, Canadian men have not seen this level of improvement over the years as nearly 10% more women complete a bachelor's degree than men (Women in Education, 2009). This seems to be due to gender-based preferences as women are more committed to school and aim to continue their education (Women in Education, 2009). According to another Statistics Canada report, 47% of male occupations are either in the fields of trades, transport and equipment operators or sales and services (Male employment, 2009). These jobs typically do not require advanced education and can explain the strong correlation in our analysis between men and having only completed a high school diploma. Men seem to be more interested in jobs that do not require education beyond high school compared to women.

Figure 8 also highlights that individuals with only a high school diploma as their highest level of education are more likely to have a greater number of children. Additionally, this also means that more educated individuals are less likely to have children. This relationship can be further justified using the work of Cohen et al., who investigated how childbearing affects educational attainment (Cohen, 2011). Like our analysis, they found that women with higher-level degrees have lower completed fertility (Cohen, 2011). This can be attributed to women with advanced education not wanting to compromise their career and educational ambitions with the financial, physical and mental stress associated with raising and bearing a child (Cohen, 2011). Similarly, we believe that educated men are reluctant to raise children when they could be pursuing higher education or attempting to advance their career trajectory.

An interesting relationship seen in Figure 8 is that individuals born in Canada were more likely to only have high school as their highest level of education. This contrasts with those not born in Canada who are more likely to have advanced post-secondary education. This relationship is further exemplified in a Statistics Canada report where researchers found that 7 out of 10 immigrants had completed university-level education (Profile of internationally-educated, 2006). Whereas Canadian-born individuals only reported 41% that a university degree is their highest level of education completed (Profile of internationally-educated, 2006). An explanation for this stark contrast could be that immigrants are required to go through Canada's points-based immigration system to gain entry into the country (Six selection factors). One of the factors required to gain points is education and as a result, immigrants in Canada are likely to have the post-secondary education needed to score points in this section. Additionally, for immigrants completing their education in Canada, there is greater pressure to advance their schooling as they lack familial and financial support that a Canadian-born individual might have.

As seen in Figure 8, individuals with high school as their highest completed education are more likely to avoid marriage. A potential reason for this relationship is that individuals with only a high school diploma tend to make less money than individuals with a higher education level (Does Education Pay, 2017). As a result, these individuals might not believe they are in a financially stable enough situation to support a marriage.

When applying our results to the small and large world (McElreath, 2018) we find that we would likely yield different results. Applying our model to the small world (keeping it restricted to the people of Canada), we would find that our model would hold up quite well. The model gives fairly accurate probabilities of whether a person gets an education higher than high school for the people of Canada which makes sense because,

for the most part, people tend to want to go to university or college or are encouraged by their family to go there. This is not the case for people in other countries. In other countries, people may want to apply for their work somewhere else or can't afford university so they instead find different work after high school. This is where our model may fail in the large world because of the different levels of university enrollment compared to Canada.

Limitations and Future Work

Post-stratification is a method used in survey analysis to weigh survey results based on the total population distribution (Little, 1993). Ideally, we would weigh every variable relative to the overall Canadian population. However, we haven't weighed our data as we are limited by time and resources. In the future, researchers should perform post-stratification to weigh the variables of the survey according to the larger general Canadian population. This can be exemplified in Figures 4 and 5 from the data section. According to Statistics Canada, the male to female ratio is near 1:1 as females account for 50.4% of the Canadian population (Female Population, 2010). However, our survey results in Figure 5 show that respondents were 55% female and 45% male. Ideally, we should be weighing our data to match this general population distribution. In contrast, the results of Figure 4 are very similar to the number of people born in Canada vs not born in Canada for the general population (Focus on Geography, 2016). Therefore, the data for this variable would not be altered much after post-stratification.

To conform to the Logistic Regression model, we have converted the highest education level completed variable to a binary categorical variable (High school is the highest education completed vs High school is not the highest education completed). However, this is limiting as different forms of post-secondary education are grouped together as equals. In reality, college degrees, trades certificates and university degrees offer varying benefits and thus should be considered separately. In the future, researchers can look to perform this analysis using a different regression model that can incorporate the differences between the various post-secondary education paths.

The data used in this analysis was filtered to only include responses from individuals over the age of 23. This filter was incorporated to account for the fact that individuals between the ages of 15 – 23 might not have had the opportunity to complete education higher than high school. However, a limitation of this added process is that we are potentially removing responses of individuals, aged 15 – 23, who willingly did not continue their education past high school, have completed 2-year college programs or have completed a short trade certification.

Appendicies

Citations

- Rohan Alexander and Sam Caetano (2020). Cleaning the GSS 2017 file. gss_cleaning-1.R
- JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com.
- Australian Bureau of Statistics. Frames and Population. https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Basic+Survey+Design+-+Frames+&+Population
- Cohen JE, Kravdal Ø, Keilman N. Childbearing impeded education more than education impeded childbearing among Norwegian women. Proc Natl Acad Sci U S A. 2011 Jul 19;108(29):11830-5. doi: 10.1073/pnas.1107993108. Epub 2011 Jul 5. PMID: 21730138; PMCID: PMC3141966.

- Government of Canada. "Six selection factors – Federal Skilled Worker Program (Express Entry)". https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/express-entry/eligibility/federal-skilled-workers/six-selection-factors-federal-skilled-workers.html
- Lightfoot, Liz. The student experience — then and now. The Guardian. from https://www.theguardian.com/education/2016/jun/24/has-university-life-changed-student-experience-past-present-parents-vox-pops
- R. J. A. Little (1993) Post-Stratification: A Modeler's Perspective, Journal of the American Statistical Association, 88:423, 1001-1012, DOI: 10.1080/01621459.1993.10476368
- McElreath, R. (2018). Statistical Rethinking. Chapman and Hall/CRC. doi: 10.1201/9781315372495
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.1. https://CRAN.R-project.org/package=broom
- Daniel D. Sjoberg, Michael Curry, Margie Hannum, Karissa Whiting and Emily C. Zabor (2020). gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. R package version 1.3.5. https://CRAN.R-project.org/package=gtsummary
- Statistics Canada, General Social Survey (2017), https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501
- Statistics Canada (2017). "Does Education Pay?". https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016024/98-200-x2016024-eng.cfm
- Statistics Canada (2009). "Women in Education". https://www150.statcan.gc.ca/n1/pub/89-503-x/2010001/article/11542-eng.htm
- Statistics Canada (2009). "Male Employment, by occupation". https://www150.statcan.gc.ca/n1/pub/71-222-x/2008001/sectione/e-men-hommes-eng.htm
- Statistics Canada (2010). "Female Population". https://www150.statcan.gc.ca/n1/pub/89-503-x/2010001/article/11475-eng.htm
- Statistics Canada (2006). "Profile of internationally-educated immigrants aged 25 to 64". https://www150.statcan.gc.ca/n1/pub/81-595-m/2010084/e2-eng.htm
- Statistics Canada (2016). "Focus on Geography Series, 2016 Census". https://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-can-eng.cfm?Lang=Eng&GK=CAN&GC=01&TOPIC=7
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Claus O. Wilke (2019). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot
- Wu, Changbao, And Mary E. Thompson . Sampling Theory and Practice. Springer Nature, 2020
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Code

Code supporting this analysis can be found at https://github.com/matthewwankiewicz/sta304ps3.