Predicting Likelihood of a Male being Married using Age and Education Level

Fadlan Arif and Kar Hian

12/10/2020

Abstract

With the rates of marriage at an all time low and the rate of divorce at an all time high, this paper examines the likelihood of a person being married via they age and level of education. with the data from Canadian General Social Survey (GSS), We selected what seemed to be key factors in a person being married and based our analysis on it. Using logistic regression, we have created a model in which we can pinpoint the probability of a male to be married just by two variables. These findings help us better understand the male demographic of Canada and see where major life choices are made.

We use R Core Team (2020), Wickham (2016), Wickham et al. (2020), Wickham et al. (2019), Dowle and Srinivasan (2020), Auguie (2017) and Xie, Allaire, and Grolemund (2018).

Introduction

Throughout developed nations, rates of marriage have been on the decline (Anne Milan (2015)). This could be due to multiple factors: rates of divorce increasing (NetNewsLedger (2019)) or more personal work to focus on. This could also be due to the importance of a legal marriage decreasing, thus lowering its popularity. Basically, with how the current world is developing, new factors appear every year that compounds onto one's decision to get married. This paper focuses on the male point of view, to easier look at the data rather than mixing the two sexes. The question we are answering is: 'What is the likelihood of a male being married?'. When answering this question, we looked at the dataset and tried to focus in on key variables or what we thought were major indicators for marriage: age and education level. Our thought behind choosing age is that many people; before their first marriage; tend to set a goal of what age to be married by. With this goal, they tend to marry at a younger age, leading to higher rates of marriage for those in their 20s. We also chose education level as those who have higher education level would think more thoroughly before entering marriage. They would investigate the many pros and cons and the effort needed to be put into marriage while others may just jump into marriage with little to no hesitation. They may just jump into marriage due to being uneducated or spontaneity. We limited ourselves to these variables as to not over-complicate the possible models and results.

Our approach within this paper was to find a binary way to classify all our data. Therefore, we decided on probability of being married, as marriage is a binary variable. Two people are either married or not, there is no in-between. We turned it into a binary variable by examining the ever_married column and applying 0s and 1s to these answers. This allowed us to look at the data from a numerical standpoint. With our question being a binary one, we agreed that logistic regression was the best method to carry out our analysis with. This is because it models a binary dependent variable (state of marriage). Then, we chose our independent variables of interest (age and education level). Afterwards, we worked on cleaning the data, extracting only the columns we needed. We then created a general linear model to compute our coefficients for our logistic regression model. Since our 'education level' variable is categorical, dummy variables were created that could be turned on and off for each category. We then created the model with our given coefficients and variables. Creating several curves, each representing different education levels. Logistic regression has the following model:

$$log(\frac{p}{1-p}) = B_0 + B_1 x_1 + ... + B_k x_k \quad (1)$$

Where p is the probability that the event of interest occurring (marriage), B_0 is the y-intercept and $B_i, 1 \leq i \leq k$, coefficient represents change in log odds for one unit increase in x_i.

With this data we see a general trend from all the curves that the probability of being married increases with age. They all have a similar 'S' shape increasing more towards the higher age. the main difference between these curves is their placement on the y-axis. They seem to almost be the same curves but shifted up and down. This is from the differing levels of education. We can see that those with the lowest probability of being married, are the ones with 'Less than high school diploma or its equivalent', while the ones with the highest probability of marriage is 'University certificate, diploma or degree above the bachelor's'. This is interesting as they are on two ends of the spectrum when looking at education levels. Through analysing the data, creating models, graphs, and charts, and discussing the results, we examine the reason for such an occurrence.

Data

The data for this analysis is sourced from Canadian General social survey (GSS) through the University of Toronto library. The GSS collects information from persons aged 15 and over in 10 provinces of Canada, excluding full-time residents of institutions. The target sample size of the GSS is 20,000 respondents. The GSS uses Statistics Canada's common telephone frame, which combines landline and cellular telephone numbers from Address Register, Census of Population and various administrative sources which have been integrated with Statistics Canada's common dwelling frame.

The data is collected using a combination of self completed online questionnaires and telephone interviews(Social and Aboriginal Statistics Division (2019)). Most often income data, are drawn from tax or other administrative files rather than direct survey questions to reduce respondent burden and to improve data accuracy. The GSS also uses Age-Order method to select a respondent within a household through an invitation letter rather than the traditional household rostering method. The methods for data collection for the GSS are ever changing to keep up with the times to reduce non-response. The GSS uses a two-stage sampling design. The sampling units are the groups of telephone numbers. The final stage units are individuals within the identified households. Survey estimates will be adjusted to account for non-response cases (Social and Aboriginal Statistics Division (2016)).

The key features of the GSS it that is covers a wide variety of themes such as Canadians at Work and Home, Family, Caregiving and Care Receiving, Giving, Volunteering and Participating, Victimization, Social Identity, and Time Use. Since the survey is conducted by statistics Canada, this survey can capture the general population of Canada.

First, the raw data is cleaned using scripts provided by Rohan Alexander and Sam Caetano (2020). After that, we read in the cleaned data from gss.csv. Our variables of interest were 'ever_married', 'sex', 'age' and 'education'. From there, we mutated the data set to create binary values for the 'ever_married' data, with a 'Yes' equalling 1 and a 'No' equalling 0. This new column of data was called 'bin_evermarried'. Afterwards we created a new dataset, selecting only our variables of interest and the new 'bin_evermarried' column. We also filtered out the non-responses, leaving us with only the responses with these variables filled in.

```
##   ever_married bin_evermarried
## 1           No               0
## 2          Yes               1
## 3          Yes               1
## 4          Yes               1
## 5           No               0
## 6          Yes               1
```

Above shows the conversion from yes/no to binary values for a few responses. Turning this into binary values help us quantify these answers and plot them on a graph. This new 'bin_evermarried' column now turns into our dependent variable.

```
##      sex  age                                                    education
## 1 Female 52.7 High school diploma or a high school equivalency certificate
## 2   Male 51.1                                  Trade certificate or diploma
## 3 Female 63.6                 Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
## 4 Female 80.0 High school diploma or a high school equivalency certificate
## 5   Male 28.0  College, CEGEP or other non-university certificate or di...
## 6 Female 63.0 High school diploma or a high school equivalency certificate
```

The data frame above shows the first few entries of our variables of interest; age and education. We filtered out the female respondents later in the process. The plotting of this data is done later in the results section.

Model

After molding the data to our needs, we compute the generalised linear model with bin_evermarried as the dependent variable and sex, age and education being our independent variables.

We then called onto the summary(first_logit) function to retrieve all the needed coefficients and assigned simpler variable names to each value to form our regression formula:

- Intercept = -2.669
- sexMale = -0.247
- age = 0.079
- educationCollege = -0.069
- educationHighSchool = -0.324
- educationLessThanHighSchool = -0.669
- educationTrade = -0.207
- educationUniversity = -0.022
- educationLessThanUniversity = 0.125

Then after using equation (1):

$$\begin{gathered} log(\frac{p}{1-p}) = Intercept + sexMale(x_1) + age(x_2) + educationCollege(x_3) + \\ educationHighSchool(x_4) + educationLessThanHighSchool(x_5) + educationTrade(x_6) + \\ educationUniversity(x_7) + educationLessThanUniversity(x_8) \end{gathered} \tag{2}$$

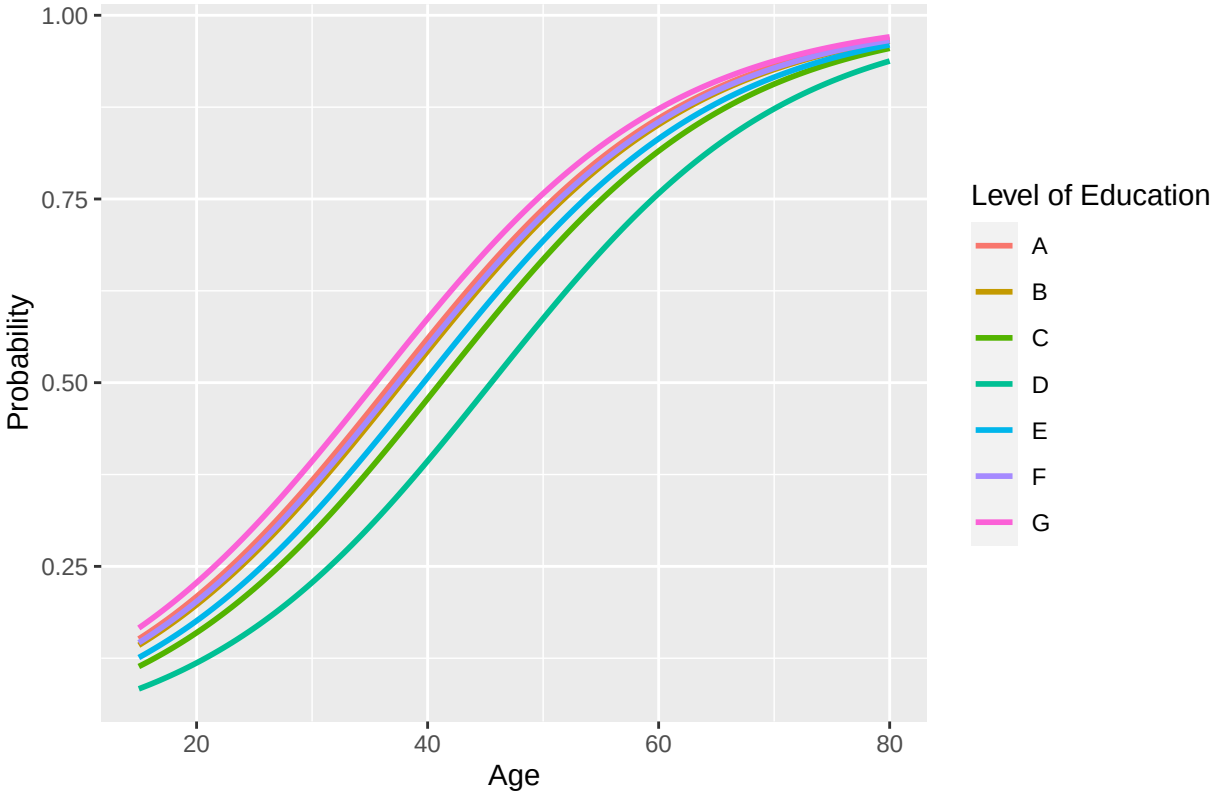
Then when placing the coefficients in:

$$log(\frac{p}{1-p}) = -2.669 - 0.247x_1 + 0.079x_2 - 0.069x_3 - 0.324x_4 - 0.669x_5 - 0.207x_6 - 0.022x_7 + 0.125x_8 \tag{3}$$

When modeling the equations, x_1 is always 'on'($x_1 = 1$) since we are looking into only males, x_2 is the age and $x_3 - x_8$ are all dummy variables. This means that when one is active (the corresponding $x = 1$), the others are all 0. When looking for our first curve for the education level: 'Bachelor's degree (e.g. B.A., B.Sc., LL.B.)', we set $x_3 - x_8$ to 0 as they represent the other six levels education.

The model below was created by separately calculating the curves for each level of education. The probability being the dependent variable and the age being the independent variables we needed to create 7 separate graphs to take into consideration the third variable: education level. We then plotted each graph onto one plane to be able to compare.

Figure 1: Probability of being Married by Age and Education Level



Levels of Education:

- **A**: Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
- **B**: College, CEGEP or other non-university certificate or diploma
- **C**: High school diploma or a high school equivalency certificate
- **D**: Less than high school diploma or its equivalent
- **E**: Trade certificate or diploma
- **F**: University certificate or diploma below the bachelor's level
- **G**: University certificate, diploma or degree above the bachelor's

Results

Figure 2: Sex distribution of respondents

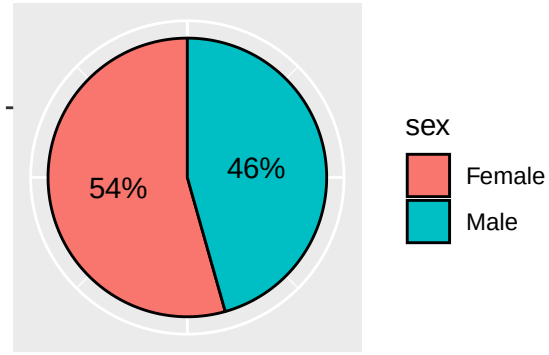


Figure 3: Age groups of respondents

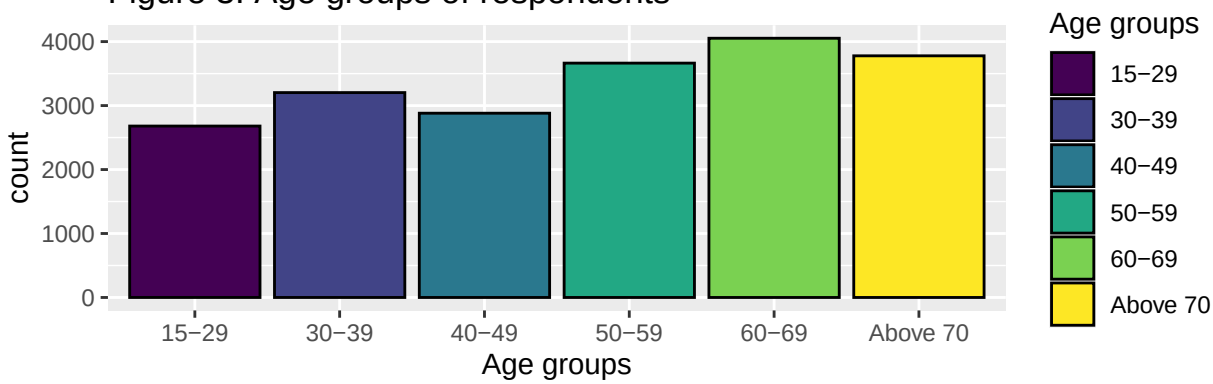
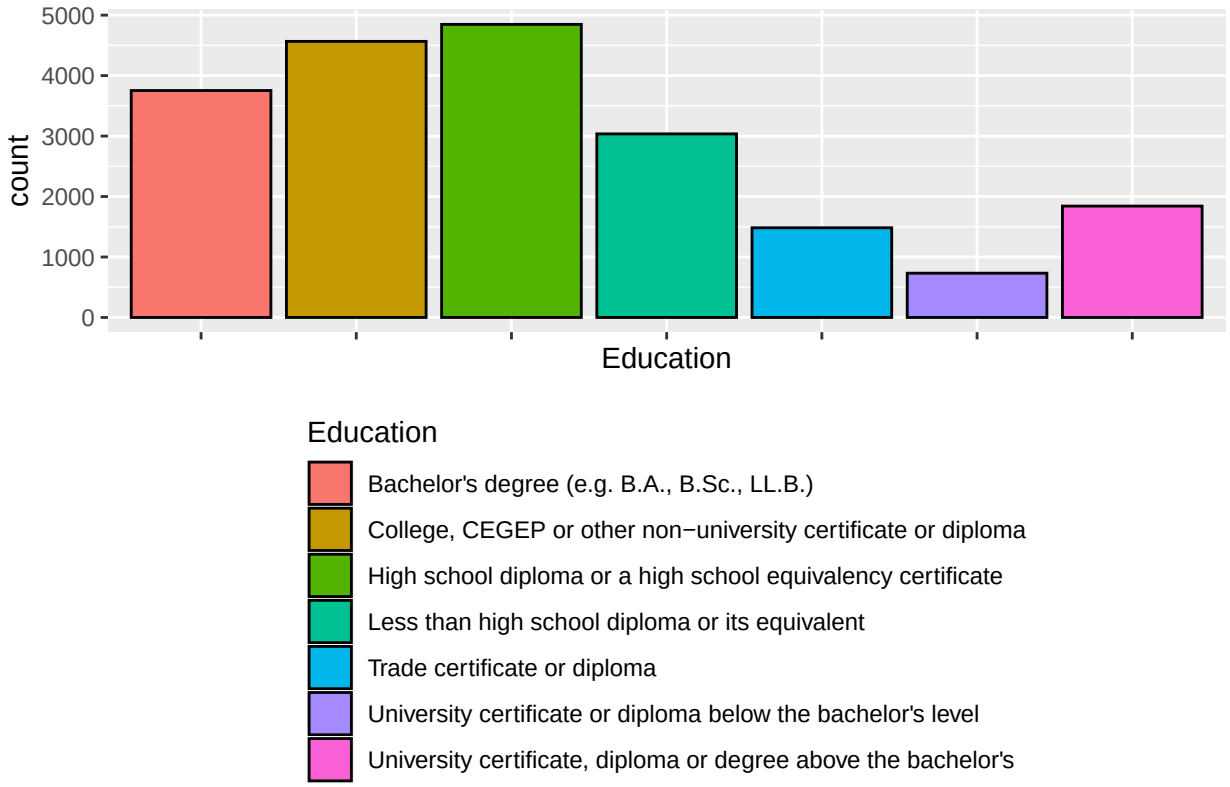


Figure 4: Level of education of respondents



Discussion

From examining Figure 4, we see that most of the respondents have education levels below the bachelor's degree. There are less people on the other end of the spectrum with a university certificate or above a bachelor's degree. This distribution of education levels helps us better understand our regression model and why it looks that way it is.

Looking at the Figure 1, we see the obvious trend that is, the older anyone is, the more likely they are to be married. This is a logical finding as people all over the world see marriage as a mandatory milestone in life and some even see it as the goal in life: settling down, having children, etc. So by logic, people will try to get married sometime in their life thus there being a higher probability of someone being married at an older age.

Another trend also persists in this model, which may be hard to spot. That is, the higher the level of education for the male, the higher probability they are to be married through out all ages. We see that line **G** (University certificate, diploma or degree above the bachelor's) at the top, followed by **A** (Bachelor's degree (e.g. B.A., B.Sc., LL.B.)), then **F** (University certificate or diploma below the bachelor's level), then **E** (Trade certificate or diploma), then **C** (High school diploma or a high school equivalency certificate), and **D** (Less than high school diploma or its equivalent). There are multiple reasons for this occurrence. One reason is that those with higher level of education, tend to have higher income and with that higher income, would be able to support marriage and everything that comes with it. While those with lower education levels such as no high school degree struggle to find jobs with a steady stream of income. This leads to instability within their lives and with this they tend to hesitate to go into marriage.

Another possibility for the difference in the curves is how towards the higher level of education, less people were sampled, or maybe of the people that were sampled, less had higher education. Whatever the case is, this led to a lower variety of responses from these education levels and maybe leading to a skewed sample of what the actual probabilities are due to a small sample size.

Looking at figure 3, we can see that the ages of respondents starts at the age of 15. This is due to the GSS collects responses from ages 15 and above. Responses from ages below 15 might have been removed from dataset before publishing the survey results. Removing responses from those under the age of 15 will not affect our investigation as the minimum age of marriage in Canada is 16 years old(Wikipedia contributors (2020)).

Weaknesses

Possible weaknesses in this paper lie within the scope of our data set. Inaccuracy could occur due to only taking two variables into consideration while many others were left out. This was done as a measure to lower complexity but could also be a point of weakness as more variables would increase its accuracy. This could be resolved by creating more models but replacing the education variable with another to compare differences.

Several weaknesses on data collection of the GSS is that the GSS uses Age-Order method over traditional household rostering method. Based on a research about alternative methods for the random selection of a respondent within a household for online surveys (Geneviève Vézina,Pierre Caron (2017)), this method has some drawbacks on the accuracy of the data. The source of inaccuracy come from the survey being filled up by the intended person. One way to mitigate is to display the condition in the online poll and make respondent verify that they are the intended person to fill up that questionnaire. Besides that, although the GSS have changed its method of data collection, but its database is restricted to those with telephone and address with statistics Canada. Those individuals that is not in that database will be missed having a chance to be sampled.

As for a critique of the survey itself, the length of it could be the cause for many non-responses. This is because when given a questionnaire, the longer it is the more likely the average person just wants to get it over with, thus leaving non-responses for questions that do not interest them. This leads to omission of important data that may have been filled in if the structure of the survey was changed. This could be done by making a shorter survey, with questions that cover a larger scope, or maybe separating the survey into multiple surveys as to not wear out the respondents.

Note

```
# Code and data supporting this analysis is available at: https://github.com/karhian/GSS
```

References

Anne Milan. 2015. *Marital Status: Overview, 2011.* Statistics Canada. https://www150.statcan.gc.ca/n1/pub/91-209-x/2013001/article/11788-eng.htm.

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics.* https://CRAN.R-project.org/package=gridExtra.

Dowle, Matt, and Arun Srinivasan. 2020. *Data.table: Extension of 'Data.frame'.* https://CRAN.R-project.org/package=data.table.

Geneviève Vézina,Pierre Caron. 2017. *Comparing Alternative Methods for the Random Selection of a Respondent Within a Household for Online Surveys.* Statistics Canada. http://www.asasrms.org/Proceedings/y2017/files/593998.pdf.

NetNewsLedger. 2019. *Highest and Lowest Canadian Divorce Rates by Profession.* NetNewsLedger. http://www.netnewsledger.com/2019/06/20/highest-and-lowest-canadian-divorce-rates-by-profession/#:~:text=According%20to%20M.J.%20O'Nions,divorce%20rate%20for%20newer%20marriages.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rohan Alexander and Sam Caetano. 2020. *Gss-Cleaning*. University of Toronto. https://github.com/karhian/GSS/blob/main/gss_cleaning-1.R.

Social and Aboriginal Statistics Division. 2016. *General Social Survey : Canadians at Work and Home (Gss)*. Statistics Canada. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=302914.

———. 2019. *General Social Survey: An Overview, 2019*. Statistics Canada. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm.

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, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wikipedia contributors. 2020. *Marriage in Canada*. https://en.wikipedia.org/wiki/Marriage_in_Canada.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.