GSS Analysis

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 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)

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. Baasicaly, 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, 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. Out 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 into marriage. They would look into 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 on marriage or spontaneity. we limited ourselves tho 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. This is why 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 examinining the ever_married column and applying 0s and 1s to these asnwers. This allowed us to looked at the data from a numerical stand point. 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 after choosing or independent variables of interest (age and education level), we worked on cleaning the data extracting only the columns we need. Afterwards we 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. Logisitic regression has the following model:

$$log(\frac{p}{1-p}) = B_0 + B_1x_1 + ... + B_kx_k \tag{1}$$

Where p is the probability that the event of interest occuring (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 increaing 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 occurance.

Data

The data for this analysis is sourced from Canadian General social survey (GSS) through University of Toronto library. The GSS collects infomation 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 administarive sources and which has been integrated with Statistics Canada's common dwelling frame.

The data is collected using a combination of self completed online questionnairs 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 in order 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 is 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 is that is covers a wide variaty 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 is able to capture the general population of Canada.

We first 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 values. 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.

From there, we compute the generalised linear model with bin_evermarried as the dependent variable and sex, age and education being our independent variables.

```
first_logit <- glm(bin_evermarried ~ sex + age + education, data = focused_data,
                   na.action="na.exclude", family = "binomial")

b0 <- first_logit$coef[1] #intercept
sexMale <- first_logit$coef[2]
age  <- first_logit$coef[3]
educationCollege  <- first_logit$coef[4]
educationHighSchool  <- first_logit$coef[5]
educationLessThanHighSchool  <- first_logit$coef[6]
educationTrade  <- first_logit$coef[7]
educationUniversity <- first_logit$coef[8]
educationLessThanUniversity  <- first_logit$coef[9]
```

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

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

Then when placing the coefficients in:

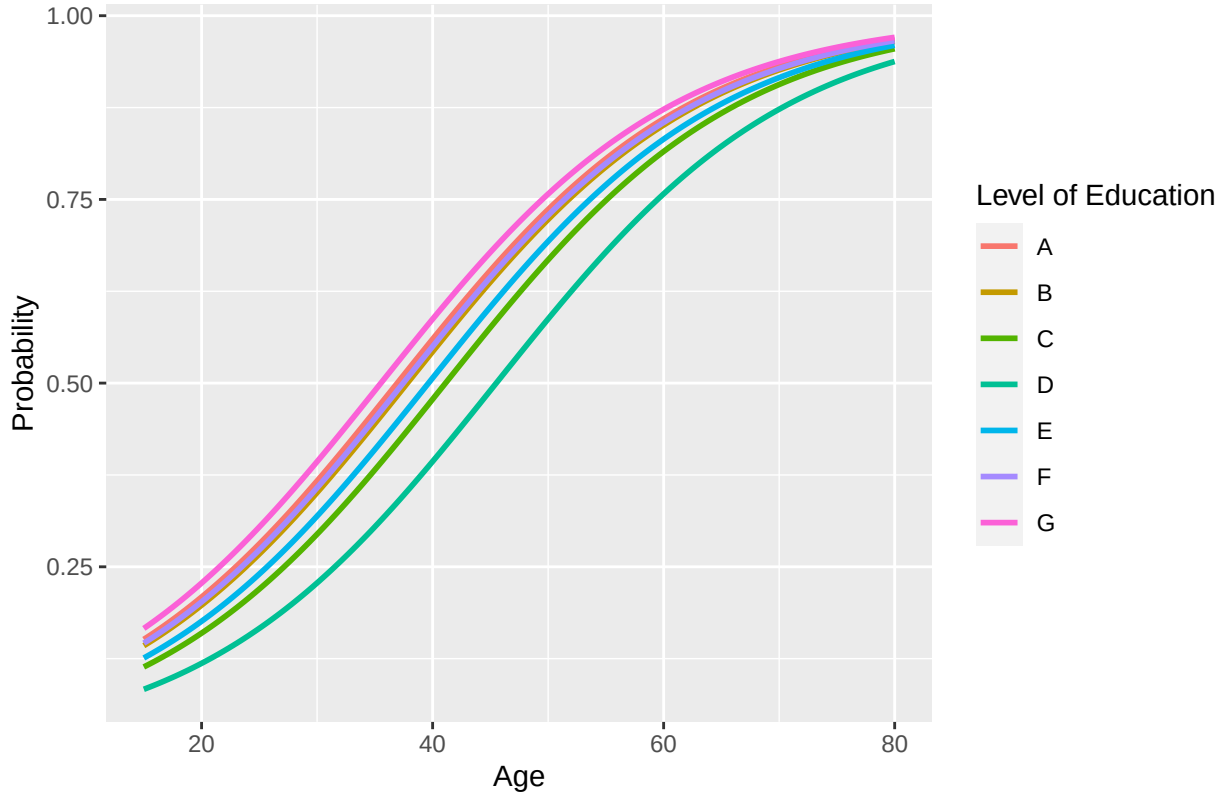
$$log(\frac{p}{1-p}) = -2.669 - 0.247x_1 + 0.079x_2 - 0.069x_3 - 0.324x_4 - 0.669x_5 - 0.207x_6 - 0.022x_7 + 0.125x_8 \quad (3)$$

When modeeling 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 mean 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. Below we show how we constrcuted the model from the given equation.

Model

The model below was created by sepearately 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 seperate 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

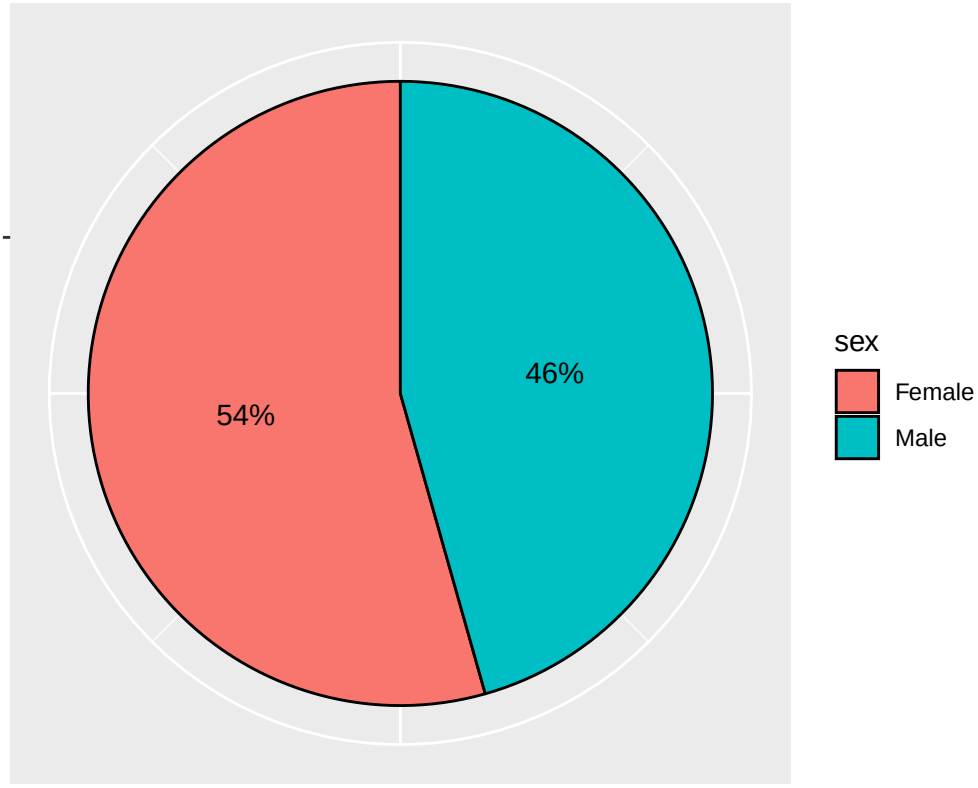


Figure 3: Age groups of respondents

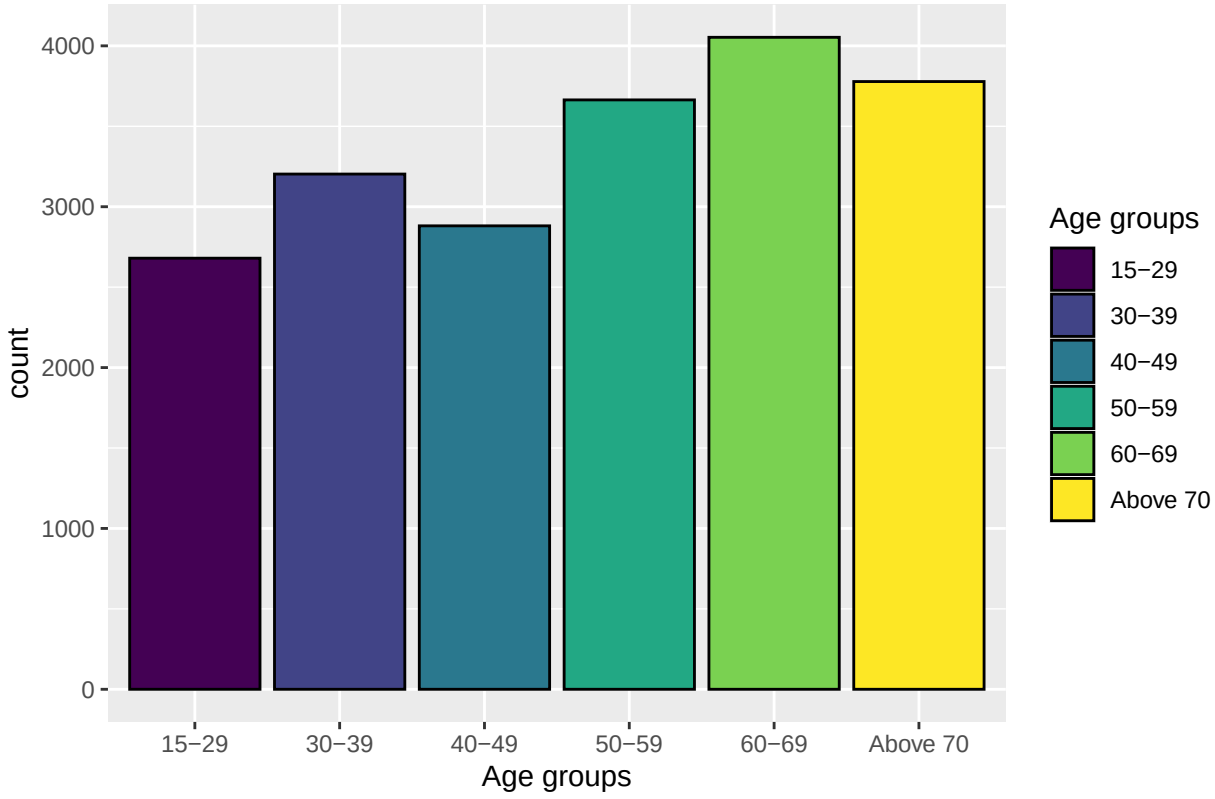
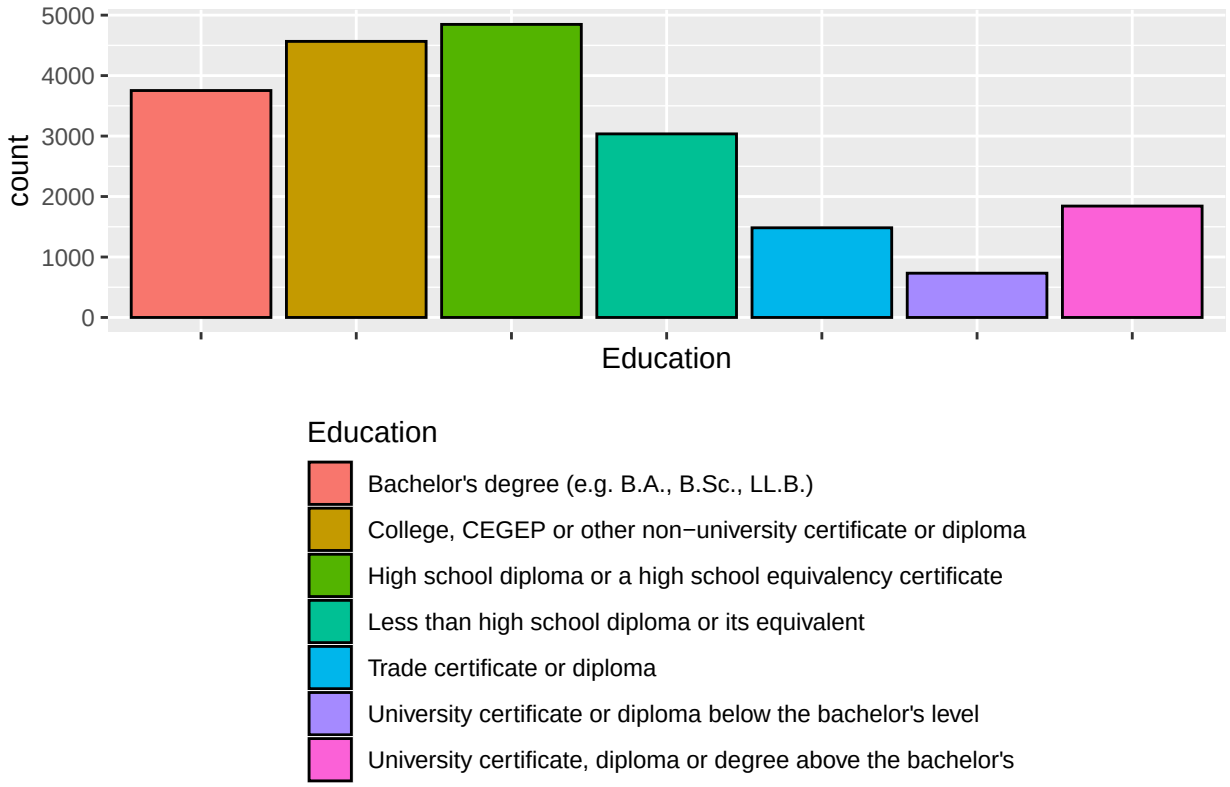


Figure 4: Level of education of respondents



Discussion

Looking at the Figure one, 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 ultimate goal in life: settling down, having children, etc.. So by logic, peopl 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), 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 occurence. One reason is that those with higher level of education, tend to have higher income adn with that higher income, would be able to support marriage and everything that comew 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 going into marriage.

points to discuss -talk about seperation between probabilities for the diff education levels - look and sex difference - look at different levels of education

Weaknesses

Possible weaknesses in this paper lie within the scope of our data set. Inaccuracy could occur sue to only taking two variables into consideration while many others werre left out. This was done as a measure to

lower complexity but could also be seen as a point of weakness as more variables would increase its accuracy. The GSS uses Age-Order method over traditional household rostering method. Based on a reseach about alternative methods for the random selection of a respondent within a household for online surveys (Geneviève Vézina,Pierre Caron (2017)), this method have some drawbacks on the accuracy of the data.

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.

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/.

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.