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. This could be due to multiple factors: rates of divorce increasing 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.

with this data we see aa 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

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.

```
raw_data <- read.csv("gss.csv")
raw_data <- raw_data %>% mutate(bin_evermarried = if_else(ever_married == "Yes", 1, 0))
focused_data <- raw_data %>% select("ever_married", "bin_evermarried", "sex", "age", "education") %>% fi
show_married <- focused_data %>% select(ever_married, bin_evermarried)
head(show_married)
```

```
##   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. 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

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

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

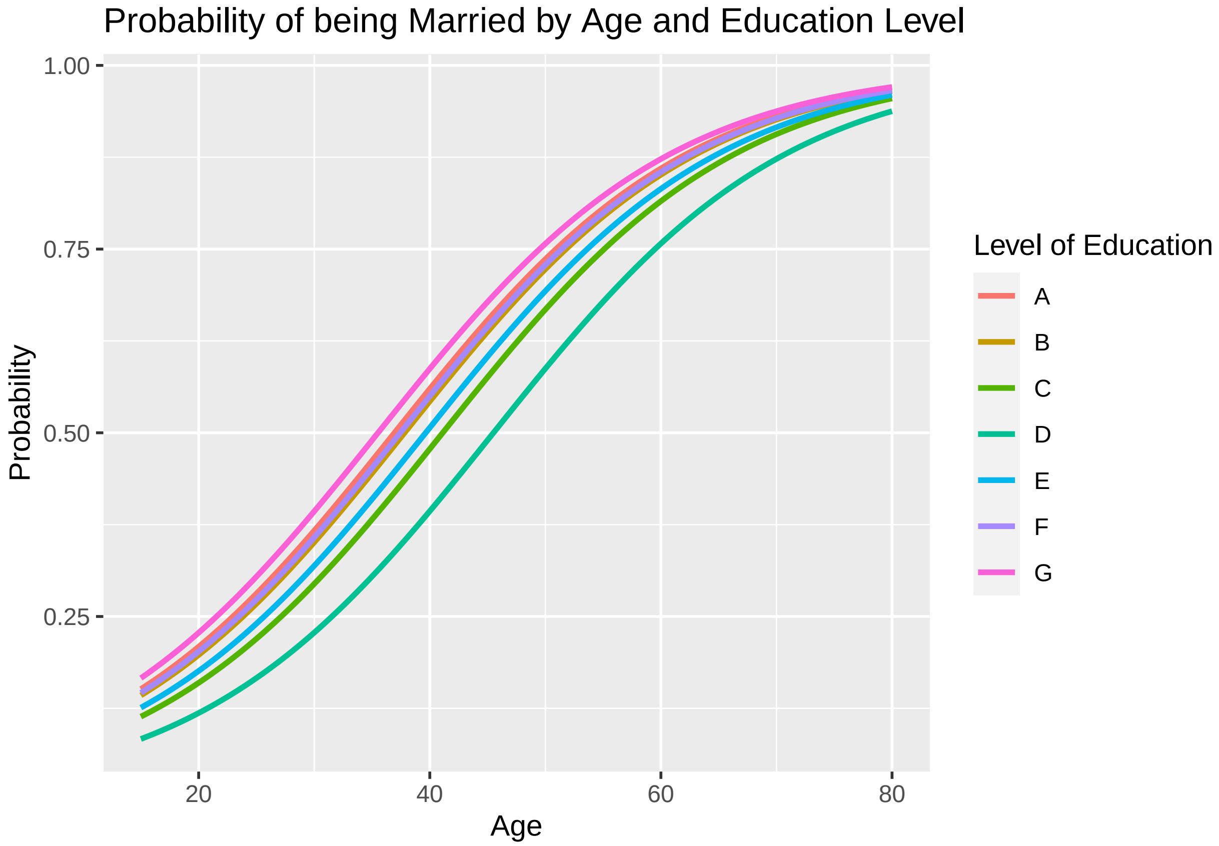
Then when placing the coefficients in:

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

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

Below is graph.



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

Discussion

Weaknesses

References

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.