Marital Status Analysis Using Logistic Regression Model

Jingyao Wang, Zhen Xia, Ruolan Zhang

October 19th 2020

Abstract

Increases in population and the improvement in society have contributed to apparent changes in the marital status, hence it is necessary for us to investigate what factors matter and how these impact the public. We use cross-sectional data provided by the 2017 Canadian General Social Survey (GSS) and build logistic regression models to identify which factors were associated with marriage status for citizens over 15 years of age. The demographic characteristics of marriage status changes with age, sexuality, family income, total numbers of children, feelings of life and individual education level. The group of people in marriage more tends to be young, male, having more children, feeling better for life while more educated and higher-income groups are the opposite.

Introduction

In 2019, about 30 million are of marriageable age in Canada's 37 million population. The proportion of the married population is declining over time. For example, in 2011, 46.4 percent of people over 15 were legally married, while 53.6 percent were unmarried (i.e., never married, divorced, separated, or widowed). By comparison, in 1981, 60.9 percent of people over 15 were married while only 39.1 percent were unmarried. As Canada's population grows, marriage rates are influenced by many factors, such as gender, education, and income level, etc.

According to the latest data by 2019, there are more married men than women in Canada, with the data 7.14 million married men and 7 million married women. Given that there are fewer men than women in the Canadian population demographics of the same year. It is investigated that Canadian men marry more than women.

Other national surveys demonstrate a positive relationship between education and marriage rates in North America by 2006. Those with college degrees are more likely to get married than those with less education. At the same time, the percentage of university degree holders living in common-law relationships, which is 12 percent, is much lower than those of low-education level people, which is 8 percent.

Over the past 40 years, the married has had higher incomes than the unmarried. Among married adults at all levels of education, men's household incomes rose more than women's. Higher real incomes help stabilize families and reduce divorce rates. According to the American community survey (ACS) 2012-2016 five-year estimate, poverty and housing are associated with the marriage of young people aged 18 to 34.

Average annual wages and homeownership for all types of workers are both associated with higher marriage rates. The economic characteristics of both genders are closely related to the marriage of today's young people.

Data

The target population includes all non-institutionalized persons whose ages are equal or larger than 15 years old in 10 provinces of Canada. Residents of the Yukon, Northwest Territories, and Nunavut were excluded, as well as full-time residents of institutions. The target frame combines a list of telephone numbers in use with lists of all dwellings within the ten provinces(AR). The target sample size is 20,000, but actually 20,602 households were chosen from the frame to receive the telephone interview. The sampling strategy called stratification was applied. Each of the ten provinces are divided into strata based on geographic areas, then Simple random sampling was performed in each stratum. For each household, a random respondent was selected to take part in the telephone interview, that forms one observation in the dataset. If respondents were reluctant to take part in the interview, they will get another 2-3 extra phone calls to explain the importance of the survey in order to encourage them to complete the interview. If nobody was at home, they would receive the call at another time period. Generally, non-responding telephone numbers can be divided into 3 groups: partial non-response, non-response with auxiliary information from sources available to Statistics Canada, and complete non-response. These non-responding telephone numbers will be directly dropped, and the weight was shifted to the responding telephone number.

The questionnaire was designed to investigate 14 sub-sections. Several equations allow for write-in questions, which are open-ended questions. They are able to collect rich responses from respondents and remind researchers with some choices they did not consider before. From 2017, this survey does not ask for respondent's income directly, which diminishes respondent's burden and increases data quality both in terms of accuracy and in response rate. On the other hand, this questionnaire was conveyed through the telephone interview, which was time-consuming.

Responses were collected by the 2017 General Social Survey (GSS), which conducted from February 2nd to November 30th 2017. This dataset was obtained from Computing in the Humanities and Social Sciences (CHASS) website at the University of Toronto. Raw data contains 81 variables with 20,602 observations in total. Some key variables such as marital status, education level, income, number of children and feelings of life were clearly displayed in the dataset. The strength of this data is, these variables are related to diverse aspects of Canadian families' living conditions. They are able to conduct many researches not only on conjugal relation, but also families' socio-economic conditions, different stages of a typical Canadian family and so on. However, households without telephones were automatically excluded from this survey population, non-responding telephone numbers will be finally dropped. The consequence is, these data may lose accuracy. Besides, this telephone interview was only conducted within Canada. The conclusion drawn from this dataset may not be that representation in other countries.

The focus of this report is to analyze how some specific family factors are correlated to marital status of people in Canada. Before building a logistic model, the raw dataset was cleaned and six key variables, 'age', 'sex', 'income_family', 'total_children', 'feelings_life' and 'education' were selected among 81 variables. Only 'age', 'feelings_life' and 'total_children' are numerical variables whereas the rest of them are all categorical. These variables were chosen since by intuition, people's marital status have relatively strong connection with these factors compared to other provided variables in the dataset. Variables such as 'age_of_first_child', 'hh_size' and 'self_rated_health' were not selected as key features. Not only becasue they rarely have direct relationship with marital status, but also some of them have too much missing values. In order to prepare a proper dataset for building a model, observations with missing value were all removed. At this time, there were 13,388 observations left, which were the sample for the model. A new variable called 'marry' was created, any respondent whose marital status was married or living common-law will produce TRUE in this variable. Otherwise, FALSE will be returned.

Model

In the model section, start to build a logistic model to predict the result of marital_status using 6 factors.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} \quad (1)$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6+\beta_7 x_7+\beta_8 x_8+\beta_9 x_9+\beta_{10} x_{10}+\beta_{11} x_{11}+\beta_{12} x_{12}+\beta_{13} x_{13}} \quad (2)$$

p: the probability of a respondent's marital status was one of "Living common-law" or "Married".

Numerical variables:

x_1 is the age of respondent.

x_7 is the total number of children in the respondent's family.

x_8 is the score of respondent's feeling towards life.

Categorical variables:

x_2 is the gender of respondent.

x_3 to x_6 are dummy variables,

$$x_3 = \begin{cases} 1 & if\ family\ income\ is\ \$25,000\ to\ \$49,999 \\ 0 & otherwise \end{cases}$$

$$x_4 = \begin{cases} 1 & if\ family\ income\ is\ \$50,000\ to\ \$74,999 \\ 0 & otherwise \end{cases}$$

$$x_5 = \begin{cases} 1 & if\ family\ income\ is\ \$75,000\ to\ \$99,999 \\ 0 & otherwise \end{cases}$$

$$x_6 = \begin{cases} 1 & if\ family\ income\ is\ less\ than\ \$25,000 \\ 0 & otherwise \end{cases}$$

x_{10} to x_{13} are dummy variables,

$$x_{10} = \begin{cases} 1 & if\ education\ is\ high\ school\ diploma\ or\ a\ high\ school\ equivalency\ certificate \\ 0 & otherwise \end{cases}$$

$$x_{11} = \begin{cases} 1 & if\ education\ is\ less\ than\ high\ school\ diploma\ or\ its\ equivalency \\ 0 & otherwise \end{cases}$$

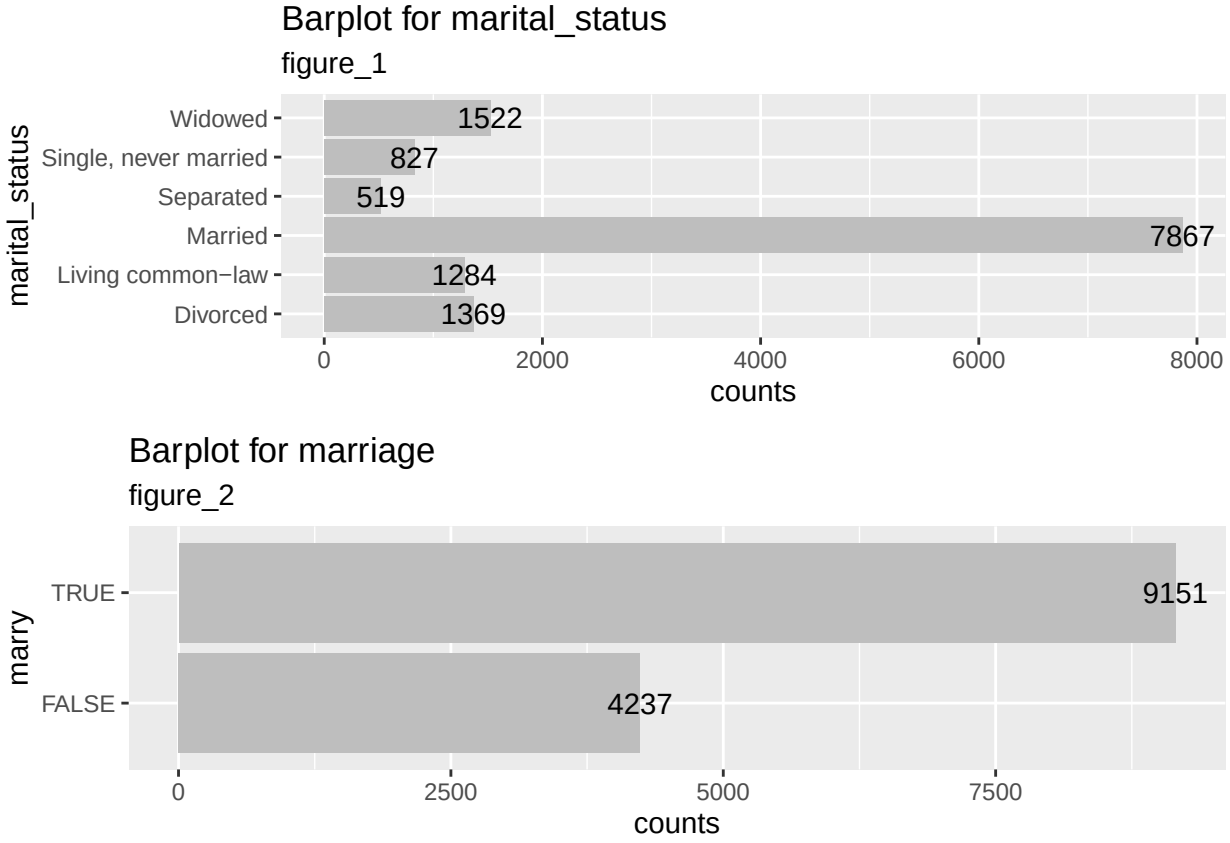
$$x_{12} = \begin{cases} 1 & if\ education\ is\ trade\ certificate\ or\ diploma \\ 0 & otherwise \end{cases}$$

$$x_{13} = \begin{cases} 1 & if\ education\ is\ university\ certificate\ or\ diploma\ below\ the\ bachelor's\ level \\ 0 & otherwise \end{cases}$$

Result

Result of Data

The number of people in each marital status were provided in Figure 1. Figure 2 combines 'Married' and 'Living common-law' into a 'marriage' group, and the rest 4 status were classified as 'unmarried'. In total, 13,388 observation are provided. In Table 1, values for each β are stated as below.



Interpretation of Figure 1 and Figure 2

Bar plot for marital_status(Figure 1) compares the number of people in six different marital status in Canada over the year of 2017. The width of each bar represents the measured value of each group of people. There are 13,388 observations in total. More than 50% of respondents are married, which excesses the other 5 status significantly. Respondents in separation status take the smallest part, with only 519 observations. In order to better compare the relationship between spouses, 'married' and 'living common-law' are combined as marriage, whereas the rest 4 categories are classified together. Any observation which is 'married' or 'living common-law' will produce a True value, observations other than these two will return False value. The 'False' suggests that the respondent does not have a spouse who is living with him or her. In Figure 2, two large classifications that mentioned above can be observed clearly. The number of respondents who are in their marriage is 9151, while the number of people in another classification is less than a half of 9151 observations. In consequence, around $\frac{2}{3}$ of respondents were in a harmonious relationship with their spouses.

Result of Logistic Model

Interpretation of β_k $(k = 0, 1, ..., 13)$:

Table 1: Regression Results

Variables	Coefficients	P_value
(intercept)	1.1876391	< 2e-16 ***
age	-0.0223732	< 2e-16 ***
sexMale	0.6587719	< 2e-16 ***
income_family $25,000 to 49,999	-2.1891702	< 2e-16 ***
income_family $50,000 to 74,999	-1.3845629	< 2e-16 ***
income_family $75,000 to 99,999	-0.6937417	< 2e-16 ***
income_family Less than $25,000	-3.7568023	< 2e-16 ***
total_children	0.0465365	< 8.24e-05 ***
feelings_life	0.2216741	< 2e-16 ***
educationHigh school diploma or a high school equivalency certificate	0.1211899	< 0.010540 *
educationLess than high school diploma or its equivalent	0.3122420	5.26e-09 ***
educationTrade certificate or diploma	0.0999953	0.103679
educationUniversity certificate or diploma below the bachelor's level	0.0587222	0.447617
educationUniversity certificate, diploma or degree above the bachelor's level	-0.2367231	0.000205 ***

β_0: When x_1, x_7, x_8=0, sex=Female, income_family=$125,000 and more, education=college, CEGEP or other non-university certificate or diploma, odds of marriage is equal to exp(1.18763).

β_1: As age increases by 1, log odds of marriage will decrease by 0.022373

β_2: When x_2 = male, odds of marriage is 0.6587719

β_3 to β_6 are dummy variables of family income:

baseline = income_ family = $125,000 and more

β_3: When baseline changes to income_family= $25,000 to $49,999, odds of marriage will decrease by 2.189170

β_4: When baseline changes to income_family= $50,000 to $74,999, odds of marriage will decrease by 1.3845629

β_5: When baseline changes to income_family= $75,000 to $99,999, odds of marriage will decrease by 0.6937417

β_6: When baseline changes to income_family=less than $25,000, odds of marriage will decrease by 3.7568023

β_7: As the total number of children increased by 1, log odds of marriage will increase by 0.0465365.

β_8: As the score for feelings of life increased by 1, log odds of marriage will increase by 0.2216741.

β_9 to β_{13} are dummy variables of education:

baseline = education = college, CEGEP or other non-university certificate or diploma

β_9: When baseline changes to education = high school diploma or a high school equivalency certificate, odds of marriage will increase by 0.1211899.

β_{10}: When baseline changes to education = less than high school diploma or its equivalent, odds of marriage will increase by 0.3122420.

β_{11}: When baseline changes to education = trade certificate or diploma, odds of marriage will increase by 0.0999953.

β_{12}: When baseline changes to education = university certificate or diploma below the bachelor's level, odds of marriage will increase by 0.0587222.

β_{13}: When baseline changes to education = university certificate, diploma or degree above the bachelor's level, odds of marriage will decrease by 0.2367231.

Discussion

Analyses of the 2017 General Social Survey (GSS) outstand important features of a person who owns marriage. The coefficient of the age term is -0.022373, which means young people are more likely to be in a harmonious relationship compared to older people. People in relatively smaller ages always play an active role in the relationship, they are much more passionate about finding love or enjoying love than large-aged people. In addition, for people who already get married for several years, it's hard to say that everybody can always remain enthusiastic towards their spouses.The loss of affection causes many couples to live separately or even lose their marriages.

The study also showed that men are more likely to be in a good marriage than women based on the β value 0.6587719, which is positive. Genetically, males pretend to be more rational than females in a relationship, they are able to adjust themselves when confronting problems in marriages. However, females are much more emotional, any small quarrel could decrease their satisfaction level about marriage.

Another supporting feature is the 'Feelings of Life' indicator which is provided by the Survey. For respondents who vote a high score for Feeling of Life, they tend to share a relatively stable marriage relationship. Naturally, Humans have a strong desire for companionship. It suggests that people who are in stable relationships tend to be happier in their lives. Moreover, for those unmarried people who have stable income and high happiness are more likely to gain recognition from others and close relationships.

Besides, the total number of children in a family is also an essential factor. Based on the logistic model, parents with more children seem to share a more peaceful family atmosphere. In reality, kids often work as bonds and tie the whole family together.

When considering how income of family contributes to the result, there is an obvious pattern that as the reduction of family income level, starting from the baseline to $125,000 and more, the odds of good marriages displays an incredible decline. A wealthy family is able to provide themselves with high living quality, which means that they will not be annoyed by any money-relevant issues. Consequently, a harmonious family relationship can be built under this circumstance.

Surprisingly, the higher the education is, the lower the odds of participating into an expected marriage. Compared to the baseline in the logistic regression model, people who only received less than a high school diploma are likely to get the highest odds, whereas those with a degree higher than bachelor's receive the lowest odds. People with a lower educational background might be easily satisfied with a simple lifestyle. They only hold basic demands for daily life, which implies that happiness can be easily obtained than those who own higher diplomas.

Among the formerly mentioned factors, education performs the most obvious influence and the factor that has the greatest impact on marriage status. Also, it is easy for governments to imply relevant policies. Therefore, the government may provide some compulsory education resources or free studying opportunities for people from any age and classes. The policy can not only achieve the goal of improving the general quality of the public hence increase marriage and fertility rates, but also bring further benefits such as raising the average income level and social stability.

In the logistic regression model, most factors can influence each other, and the logic is self-consistent. People who have close relationships tend to have children and usually enjoy more about their lives. Those with higher education can also earn more, hence they have higher demands on quality of life and spouses, which makes more single people. In the marriage market, well-qualified men are often more popular than women. Young are bound to be more popular than the old.

Yet the model only represents part of the large world. Of all the variables that affect marital status in Canada, there must be ones that are not included in our model. The most obvious ones are the legal and cultural background of different countries. These factors are harder to be quantified or investigated, but they can actually affect every aspect of everyone's life. Nonetheless, the logistic model we built in the small world makes enough sense.

Weakness

There are around 20,000 sample sizes in total, and the overall response rate is only 52.4%. This is not enough to draw a general conclusion. Part of the reported findings may be biased by the reliance on a telephone sampling strategy. Also, the sample may be too one-sided due to only including men and women in its choices, and this doesn't take the transgender into account. We have to satisfy its assumption of the logistic model when it is used to conduct this survey. In this case, each observation satisfies the assumption, which is, the variables are independent of each other. These assumptions would simplify the model while being troublesome if leading to multicollinearity. The above assumptions might lead to a procession error, that is, a non-sampling error. Moreover, in the logistic model, β is assumed to be a fixed value, but this β value is more reasonable to follow a certain distribution and hence be estimated.

Next Steps

In the future, we might want to use Bayer's model. The value should be considered to follow a certain distribution rather than a fixed value. We could also add more survey options because this would make a more representative sample. To give an example, put more gender options other than man and woman in the survey.

Right now, we are using SRS, alternatively, we can use Stratified Sampling in different provinces. Because of the different geographical location, tax revenue, and cultural background of different provinces, it is possible to influence the main objectives discussed by the Survey. Besides, the survey should contain more provinces except for the existing ten to represent the whole Canadian society.

References

Number Of Married People in Canada, By Gender 2000-2019 Published Duffin-Mar 12 -\ https://www.statista.com/statistics/446111/married-couples-in-canada-by-gender/

Women, Men and the New Economics Of Marriage Richard Cohn -\ https://www.pewsocialtrends.org/2010/01/19/women-men-and-the-new-economics-of-marriage/

Figure 1 Population Pyramid Of Legal Marital Status By Single Year Of Age and Sex, Canada, 1981 and 2011 Statistics Government of Canada -\ https://www150.statcan.gc.ca/n1/pub/91-209-x/2013001/article/11788/fig/fig1-eng.htm

Caetano, S. (2020). Gss [Csv]. Toronto: Samantha-Jo Caetano.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan. Boca Raton: Chapman & Hall/CRC.

(2020).Public Use Microdata File Documentation and User's Guide[PDF file].Ottawa.: authority of the Minister. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

Appendix

Code and data supporting this anaylysis is available at: https://github.com/Wang-Lucy107/Analysis-on-2017-GSS