# Factors that can influence Marital Status

WenXuan Zhai

December 2020

Code and data supporting this analysis is available at:
https://github.com/WenXuan-Zhai/Sta304Final

## 1. Abstract

First and Foremost, this report is researching on marital conditions in Canada and investigating on how external factors influence on Canadian marital decisions. It is discovered that people who have higher educational level and financial condition are more likely to being in a marriage. Understanding what most of the population care in a relationship helps policy makers in improving the society since a stable society is closely related to citizen's life quality.

### Keywords

Canada, Marriage, Logistic Regression, Observational Study, 2017 GSS dataset

## 2. Introduction

```
Statistical analysis could often reflect current social situations; peo
ple could scientifically utilize professional analysis tools to underst
and changes in people's thoughts. As diverse cultures and tolerance of
minority ideas gradually increase around the world, it is worth explori
ng and studying how human behaviours are transforming. One effective wa
y is to look at how factors are affecting marital status in the country;
 not only young people's attitudes toward marriage are shifting nowaday
s, but the living habits of married couples may also be changing. Moreo
ver, rapid development of technology is causing the gap of social roles
 between genders getting smaller.

Using observational survey data and objective information, it is practi
cal to seek influential parameters that are closely related to marital
status. To be specific, we are interested in the characteristics of mar
ried and never married populations and what factors force them to make
such a decision. Although the analysis itself cannot change human's tho
ughts and behaviour, it could show any existing social issues and refle
ct how current policy could be improved.

To dive deeper into the dataset and to cooperate with the main purpose,
 the response variable "number marriage" is reclassified into either ne
ver married or married once. Therefore, this analysis is using logistic
 regression to further study marital behaviour with a binary response v
ariable. Since each observation only falls into one of the two categori
```

```
es under the response variable, it is meaningless and inappropriate to
form a straight regression line as in a linear regression model. Theref
ore, probability of the target event is covered under a logit function
to make the regression line smoother and analytical.

According to Statistics Canada [1], marriage rate increased in recent d
ecades especially in years after 2016, but declined slightly in the las
t two years. In Canada, marital status is divided into four groups: bei
ng common law, married, separated, divorced and widowed [1]. Lately, th
ere has been a growing number of common law spouses than married couple
s in the country [2]. Younger populations are viewing traditions differ
ently and rejection of marriage might sometimes come from financial pre
ssure [3]. As a whole, it is obvious that people's perception on marria
ge is unstable, and understanding the reasons behind these phenomena no
t only helps sociologists and policy makers to improve their scheme, bu
t also helps unmarried people to recognize different opinions.

The 2017 GSS dataset is used in this analysis to perform a logistic reg
ression model. Secondly, details and construction of the dataset, distr
ibution and visualization of those information, and how the logistic mo
del processes the observations to draw comprehensive and reliable concl
usions will be introduced in the Method section. The result section wil
l provide some visualization and additional explanation to support the
conclusion drawn from the model. Finally, conclusions will be made in t
he conclusion section. The analysis will also discuss any weakness and
further researches that could be made in addition to this analysis. Ove
rall, the main goal of this analysis is to discuss the transformation o
f marital decision influenced by both personal characteristics and exte
rnal factors. The results are aimed to indicate factors that have a pro
moting trend on marriage decisions of Canadian residents and to demonst
rate a clearer direction for policy makers.
```

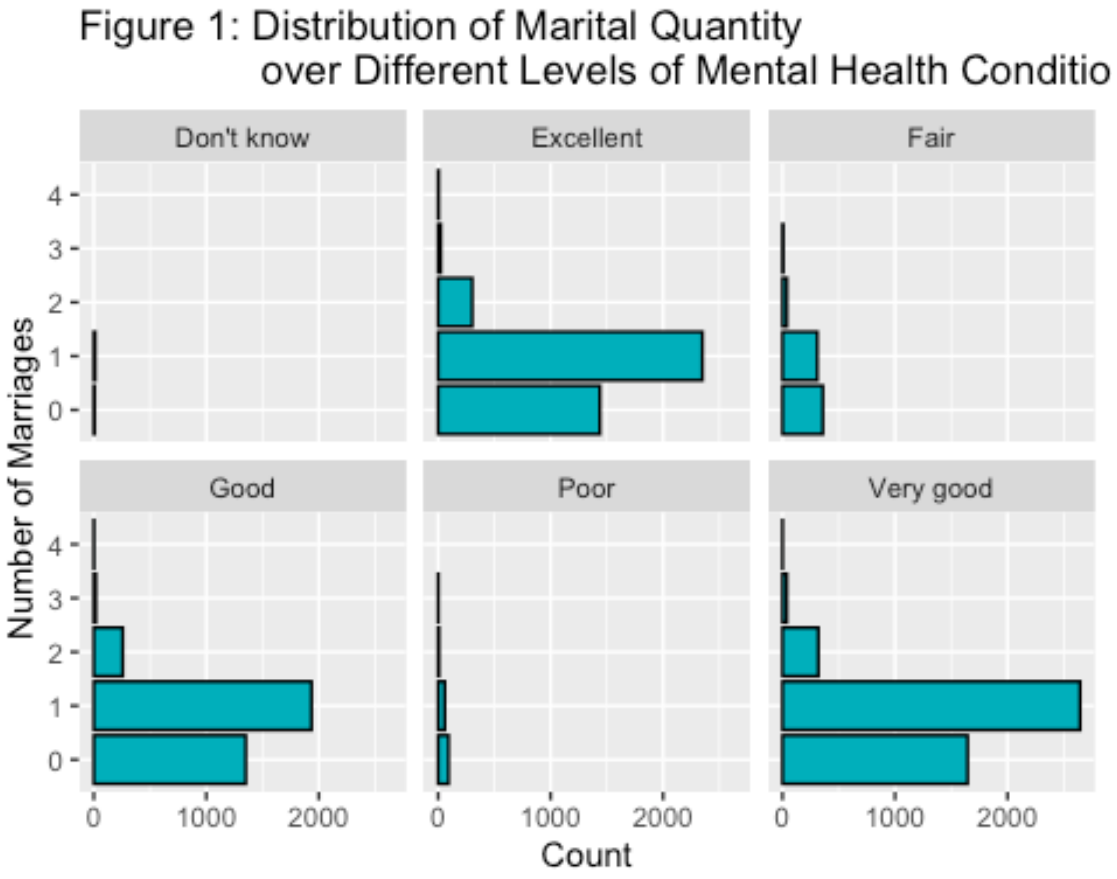
## 3. Method

### 3.1 Data

As mentioned, the 2017 GSS dataset is collected through phone calls to citizens who are 15 years and older across the 10 provinces in Canada. People who are long-term residents of collective dwellings and habitants in Yukon, Northwest Territories and Nunavut are not included in the study. The dataset is created for ensuring Canadian life quality and correcting any flaws in related policy reflected by recent social phenomena. It is using a stratified sampling technique based on several Census Metropolitan Areas (CMAs) to survey each individual in the frame population. For further sampling performance, all trained interviewers were using telephone numbers obtained from Statistics Canada and aiding with a list of residential records in Address Register (AR) to distinguish numbers from the same household. Refusal of providing a response in certain sections of this interview is not allowed. The 2017 GSS results in 20,602 observations while more than half of the respondents coordinated the investigation.

Variables involved in the dataset are mainly about personal information, family conditions and social roles. Variables that contain a lot of NA's were extracted from the data to avoid incorrect results. Also, special cases might not be recognized since not all information from the conversation in the phone call is recorded into the data, which might lead to slight inaccuracy.

```
The following figure illustrates the distribution of "number_marriages"
 over the six categories of mental health status. An interesting fact a
ddressed from this plot is that most people in their first marriage sen
se more happiness while more bachelorhood feels poor about their life i
n general. Since mental state is an important segment of daily life, it
 is closely connected to well-being of the participants and productivit
y in work. However, most unmarried people are the younger population in
 the country, they are in a stage of instability in their entire lifeti
me and may be focusing more on education and jobs. Unmarried population
s may also contain people who have just divorced as leaving a relations
hip with another human being may alter mental health.
```




### 3.2 Statistical Model

As mentioned, in order to perform a logistic linear model, the response variable is reclassified into either "never married" or "have married". Predictors selected from the 2017 GSS data could be classified into personal information which "delineate

population groups" and external factors (i.e. pop_center, education, income respondent etc.). Various statistical technique and professional standards, such as age, sex, population center, education as well as the respondent's income could all influence marriage decision. Under this binary logistic regression, probability of success in the ultimate event will be taken in a logit transformation (formula on y) followed by a regression assembled by all the significant predictors.

There is a combination of numerical and categorical variables. For example, ages could be observed in as a numerical variable which helps people to see a smooth shifts in the association. Income is discussed as categorical predictor since there is not a limit in income and it could be easily classified as middle class or higher income class etc. To be more detailed, it is categorized into more than one level instead of having only low, medium or high levels. Variables such as education and population center are also categorized in this manner. The final version of regression model is as follows, where $\beta_0$ denotes as the intercept and $\beta_1$ denotes the coefficient for the variable Age.

$$log(\frac{p}{1-p})$$
$$= \beta_0 + \beta_1 x_{Age} + \beta_2 x_{Male} + \beta_3 x_{Pop:PEI} + \beta_4 x_{Pop:RuralandSmallareas} + \beta_5 x_{Edu:College}$$
$$+ \beta_6 x_{Edu:HighSchool} + \beta_7 x_{Edu:BelowHighSchool} + \beta_8 x_{Edu:TradeCertificate}$$
$$+ \beta_9 x_{Edu:Universityabovebachelor} + \beta_{10} x_{Edu:UniversityBelowBachelor} + \beta_{11} x_{Income125000+}$$
$$+ \beta_{12} x_{Income25000-49999} + \beta_{13} x_{Income50000-74999} + \beta_{14} x_{Income75000-99999}$$
$$+ \beta_{15} x_{Income25000-} + \epsilon$$

One main guideline for selecting the final model is by using the p-value; lower p-value suggests a more significant evidence to support that the null hypothesis, the particular predictor does not have any influence on the response variable, is not true. At the beginning, a full model is created using all the predictors chosen from the dataset; however, parameters such as "self_rated_mental_health" and "average hours worked" do not show a significant sign from the 0.05 p-value. Thus, the AIC backward method and Anova chi-squared test are used to further justify the above concerns and to attain a more reliable result from the full model.

```
The step() function used for checking the final model is derived from t
he full model in a backward manner. It relatively compares models based
 on a penalty standard that tells the goodness of fit and measures whet
her the complexity of the model has assisted in providing a better regr
ession. Accounting for all potential variables and decreasing in order,
 the model with lowest AIC value will be selected through the deletion
process.
```

Anova chi-squared test is used for comparing two adjoining models. As "state of mental health" still does not perform significantly according to its p-value in the model selected from AIC, this method is used for observing accuracy of the model after extracting the one or more specific parameters. Dealing with logistic regression models, the core principle of this method is to compare correlation between variables to the dependent response. Smaller p-value under the test

indicates subtracting the variable “self_rated_mental_health” has achieved a better performance in the model.

## 4. Results

Shown in the following table, age and sex are highly premonitory predictors; one additional increase in age will multiply the odds of having married by $e^{-}2.581151$ or approximately 0.0757. For categorical predictors, the final model is suggesting that the probability of already being in a marriage is 1.086 times more for male than female while holding other predictors constant. In the same manner, more people living in rural areas or remote provinces are more likely to form their own family while the majority people living in urban cities did not. Even though the model does not show evident associations for some education and income level with marriage status, it still reveals that possibility of forming a new family is higher for high-income individuals.

Table 1: Logistic Regression Model for Marriage status with Coefficient and P-values

| | Coefficient | P-value |
|---|---|---|
| intercept | -2.58 | 0.00 |
| age | 0.08 | 0.00 |
| Sex(male) | -0.18 | 0.00 |
| Region(PEI) | 0.43 | 0.00 |
| Region(ruralsmall areas) | 0.19 | 0.00 |
| Education(College and other) | -0.07 | 0.27 |
| Education(High School) | -0.26 | 0.00 |
| Education(Less than high school) | -0.59 | 0.00 |
| Education(Trade Certificate) | -0.28 | 0.00 |
| Education(University below bachelor) | 0.06 | 0.60 |
| Education(University above bachelor | 0.06 | 0.43 |
| Income($125000+) | 0.31 | 0.02 |
| Income($25000-$49999) | -0.39 | 0.00 |
| Income($50000-$74999) | -0.11 | 0.25 |
| Income($75000-$99999) | 0.05 | 0.60 |
| Income($25000-) | -0.60 | 0.00 |

Based of the above analysis, it reveals that more male among the Canadian population have been married than female. Clearly, as the age grows, majority of the population have already married. The regression suggests that respondents living in small population centers have a significant positive effect on the response variable according to p-values of 0.000447 and 0.000521. It also hints that less education and fewer financial resources are not supportive for the participants to go into marriage from the corresponding p-values.

Since age and sex are normally immutable biological characteristic, it reflects that although modern society believes that male and female are equal in terms of human behaviour and cognition, they still have a diverse demand in marriage. People also have different mindsets at different ages, so age is importantly related to marriage decisions. At the same time, social circles and activities are highly dependent on the

population center; people living in rural areas have a limited networking environment thus might enter marriage earlier compared to those living in large urban population centers. Education and income could evidently explain a person's ability and source of survival, which indirectly alters the person's opinion.

## 5. Discussion

### 5.1 Conclusions

In the beginning, it is found that the 2017 GSS dataset was using interview questions by phone call to get the information. Stratified sampling was used based on geological regions. Not availiable observations were extracted to avoid biasness in the conclusion. Secondly, the model was built using logistic regression model aiding by the AIC backward method and Anova chi-squared test to further optimize the regression.

```
As mentioned earlier, marriage rate is found to having a downward trend
 recently and probably in future years. The Heritage Foundation [4] sta
tes that, marriage is an important sector in pertaining a stable and he
althy society but also admit that people in modern society is favouring
 more on self-reliant. Marriage is beneficial to most of the population
 that mortality rate, financial situations and mental health are more o
ptimistic with married spouses. For instance, single males are six time
s more likely to be imprisoned than married males. Men contribute more
on the community and their career are those who are in a stable relatio
nship and "living with biological children". Overall, marriage would re
sult in higher life quality to either a family or the whole society.

Finally, the regression concludes that people's vision and tendency tow
ards marriage is foreseeable. It is always revolving around several cor
e factors: financial sources, educational level, and living environment.
 Also, as age rises, people may value more on accompanying with their p
artner than living alone, supported by the very low p-value of less tha
n 2e-16. As long as reasonable legal protections could be given to both
 single citizens and married spouses, and sustain long-term supports on
 education and careers, fluctuations in marriage rate would be reasonab
le.
```

### 5.2 Weakness

Since the main goal of the investigation is focusing more on decision of weather marry or not， population who have higher number of marriages, i.e. people currently in their second or third marriage, might having a totally diverse attitude towards the incident. Moreover, observational data could only investigate facts from objective perspective but emotional and subjective factors also plays an important role in daily life. Overall, conclusions are made only on information gathered from the survey, which did not account for any specific and special circumstances. Another weakness existing in the analysis is that missing observations might also lead to slight deviations in the result and might alter in further inferences.

### 5.3 Next Steps

To improve the study and attain a better conclusion, more specific data might be needed. One important improvement might reflect in the response variable of the logistic regression. Investigations on the data could be separated for different population as discussed before. To verify applicable of the model on different datasets, the data could be separated into two sets, one training set for modeling and another testing set for validating. Afterwards, frequent updates for the dataset is also important so that the conclusions are practical for current situations. More predictors could be considered in constituting the model where recent factors are more common and may not account for any hidden problems.

## 6. References


1. Statistics Canada. Table 17-10-0060-01 Estimates of population as of July 1st, by marital status or legal marital status, age and sex
   DOI: https://doi.org/10.25318/1710006001-eng

2. Family Matters: Being common law, married, separated or divorced in Canada, Retrieved date: Dec. 8th, 2020, Statistics Canada., Date of Publication: May. 1st, 2019 https://www150.statcan.gc.ca/n1/daily-quotidien/190501/dq190501b-eng.htm

3. Arti Patel, Global News, Access Date: Dec.8th, 2020, Date of Publication: May. 7th, 2018 https://globalnews.ca/news/4191139/canadian-attitudes-marriage/

4. Marriage and Family: The Necessity of Marriage, Retrived date: Dec. 20th, 2020, The Herritage Foundation: Marriage and Family, Date of Publication:Oct. 20th, 2003 https://www.heritage.org/marriage-and-family/report/the-necessity-marriage#:~:text=The%20reason%20marriage%20is%20important,promotion%20of%20the%20common%20good.&text=Marriage%20promotes%20the%20common%20good%20by%20building%20families%20and%20raising%20children.

5. General social survey on Family (cycle 31), 2017， https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.html

6. General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User's Guide, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

7. 2017 General Social Survey: Families Cycle 31 Public Use Microdata File PUMF https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_Codebook.pdf

8. Code for access the variables present in the data framework: https://www.geeksforgeeks.org/accessing-variables-of-a-data-frame-in-r-programming-attach-and-detach-function/

9. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data, Retrieved date: Dec. 8th, 2020 https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

10. Hadley Wickham (2020). tidyr: Create tidy data, Retrieved date: Dec. 8th, 2020 https://tidyr.tidyverse.org/, https://github.com/tidyverse/tidyr/

11. Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington (2020). ggplot2: ggplot object, Retrieved date: Dec. 8th, 2020 https://ggplot2.tidyverse.org/reference/ggplot.html, https://github.com/tidyverse/ggplot2/blob/master/R/plot.r

12. Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole., Retrieved date: Dec. 8th, 2020 https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/levels

13. R-core R-core@R-project.org. step: Choose a Model by AIC in a Stepwise ALgorithm, Retrieved date: Dec. 8th, 2020 https://www.rdocumentation.org/packages/MASS/versions/7.3-53/topics/stepAIC

14. Nathaniel D. Phillips. YaRrr! The Private's Guide to R, Date of Publication: Jan. 22th, 2018 https://bookdown.org/ndphillips/YaRrr/

15. Two Way Tables. (n.d.). Retrieved date: Oct 20th, 2020, from https://www.cyclismo.org/tutorial/R/tables.html

16. Rohan Alexander and Sam Caetano, Raw data cleaning on 2017 GSS data, Publication Date: 7 October 2020, gss_cleaning.R