GSS-Report-Yi-Su

Yi Su

10/10/2020

Abstract

This report focuses on the impact of income level, age, and location on marital status. Although it is expected that marital status should not be causal with these variables, it could be shown that the logit probability of having or has a conjugal relationship is correlated with them. A Bayesian binary logistic regression model was fitted, using the 'brms' package in R, from data collected by the General Social Survey-Family which was designed and conducted by Statistics Canada in 2017. Although some constraints and limitations exist, the finding suggests that higher income levels do come with higher odd of married at least once before, and so does age. Meanwhile living in the less populated area comes with higher odd in married.

Introduction

Marriage is the recognized union of two people in a personal relationship. This personal relationship is often considered as purely emotional and not considered as something not predictable by other factors in modern society. However, this does not mean one can not assess the probability of being married/in a conjugal relationship or at least married once. In the real world, marriage may not be pure emotional, and it might be affected by other components like special situations you have been through with the person, and wealthiness of you and your spouse. In this report, we focus on assessing any potential impact on marital status due to age, annual income level, and population density in the region of residence.

First of all, the support of a real-world dataset is essential. The General Social Survey (GSS) - Family is an interview survey conducted every 5 years by Statistics Canada. This family theme survey was designed to monitor changes in Canadian families, and the information collected may show an impact on programs and policies like parental benefits. The GSS- Family-2017 dataset is a cross sectional set focusing on family information such as conjugal relationship, household status, parental history, and family annual income level, as well as general sociodemographic information like age, education, and citizen status.

In order to assess the potential impact, a Bayesian binary logistic regression model was fitted onto the GSS dataset. (Thankful to R and R package 'brms'.) Even though the results of the model are not perfect, it gave some decent directions to the potential impact. Annual Income level appears in a positive correlation with marital status and so does age. Meanwhile, the population density appears in a negative correlation with marital status. Does this answer the question? Not perfectly, but it provides more understanding of the situation. In fact, the probability of ever got married would never be modeled perfectly due to its emotional nature. In order to improve the model, the next step might be adding more predictors, and in that case, the GSS dataset does not provide more good predictors of marital status because of the nature of family theme. All detail on coding of the results can be found at: https://github.com/YiSu2000/GSS-Report/blob/master/GSS-Report-Su_Yi.Rmd

Data

The 2017 General Social Survey (GSS) - Family survey dataset targeted the population of all Canadian citizens at the age of 15 or older, but excluding residents of the Yukon, Northwest Territories, and Nunavut, and all full-time residents of institutions. Statistics Canada used a stratified simple random sampling without

replacement sample strategy. Specifically, a list of telephone numbers in use, from providers like cell phone companies and census, was linked to the address register, and each stratum was designed based on geographic location. About 86% of the phone numbers were successfully linked to an address. If an address was linked to multiple numbers, the first phone number linked, by chronological order, will be chosen. The surveying took the form of a phone call. If there are multiple participants within a household, the person picking up the call should complete the survey. To improve the response rate, numerous attempts were made to unanswered phone numbers and people who refused the first call. The target sample size was 20,000 while the actual number of respondents was 20,602, with a response rate of 52.4% which is good enough for a survey of this sample size. More detail on the methodology of the GSS survey can be found at the Statistics Canada website which is linked in the appendix.

Reliability of the truth value is one of the strengths of the GSS dataset, since it is conducted and designed by Statistics Canada, the chance of getting fraud responses should be lower than surveys conducted privately. Notice that the GSS survey is a weighted survey intended for all populations of Canada, but this feature would not be so important to the interest of this report. Another strength of the GSS is the ability to compare data from this survey cycle to a previous one. The 2017 GSS family survey is the 31st cycle, and one can compare this to the earlier cycles as a pooled cross-sectional data to assess changes of social or family behavior over time. However, this feature would not be used for this report.

The largest limitation of the GSS dataset on evaluating change over time is that the observations are not comparable at an individual level due to the protection of respondent privacy. This is not impactful to this report but might be a constraint for the evaluation of change over time.

For this report, the link to the survey question on marital status, age, annual income level, and residential population density is given in the appendix. Notice that marital status and income level are derived variable based on a series of questions. While age is a direct result start from 15 and capped at 80, all respondents at age of 80 or more will be classified as 80. The residential population density is decided based on the address register and classified as living in urban areas, rural areas, and Prince Edward Island (PEI). The questions were designed carefully to not offending the respondent and giving respondents the choice of skipping. However, there are multiple derived variables other than marital status and annual income level, meaning the number of questions might be overwhelming for respondents, and thus endangering the response rate and correctness of responses.

Data cleaning of the raw dataset from Statistics Canada was done thanks to the data cleaning code created by Rohan Alexander and Sam Caetano. To choose only the observations we are interested in, all observations under the age of thirty are excluded from the data. This filtering is to prevent the potential bias of relation between age and annual income level since the younger population does have a lower income level on average. Setting the sample population to only people above thirty should eliminate a large part of the bias. The age variable in GSS dataset is a continious variable. And annual income level is a categorical variable with six levels, with the lowest level cut at below $25,000CAD$ and highest level capped at more than $12,500CAD$ The population density indicator in GSS dataset is a categorical variable with three level. See the legend of Figure A1&A2 for details on the level of categorical variables.

For the next step, a dummy variable was created based on the marital status of each respondent, with 1 as married at least once, and 0 as never married before. Notice that living common-law is defined as married here, since living common-law is defined as a conjugal relationship by the government of Canada, and it can be considered as married on an emotional basis. The dataset is then separated into a test set and train set. The train set contains 80% of the original dataset that was randomly selected, and the test set contains the remaining 20%. This separation of the dataset is essential for cross-validation, which would help assess the quality of the regression model and is worthy to take the risk of reduced sample size. Figure A1 to A3 in the appendices provides some visualization of the test dataset.

Model

This report focuses on modeling marital status by age, annual income level, and population density around the residence. Income level and age are natural predictors one can think of when relating to marital status.

Higher age often has a higher chance of married or ever married, and a supportive income level is often the foundation of a stable and long term conjugal relationship. Population density around residence potentially affects the number of options or the chance of meeting one's future spouse. However, the expected direction of the impact of residential population density is unknown since it would be reasonable and arguable both ways.

A Bayesian binary logistic regression model was fitted on the training dataset. Binary logistic regression is the appropriate regression for the binary dummy variable on ever got married before, it models the logit probability of got married before. There are be eight explanatory variables in this model, one for age, five for the six levels of income, and two for the three levels of population density. Income level and residential population density both have one less variable than their categorical level because of the dummy variable coding. The one level without a explicit variable appears implicitly when all other variables for the variable equal 0. In this case, the interpretation of βs would be different, it now represents the estimated difference between the effect of the explicit dummy variable and the implicit dummy variable.

- The model may be expresses as:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{i=1}^{8} \beta_i x_i$$

- With Prior, for j from 0 to 8:

$$\beta_j \sim Normal(0, 10)$$

A global non-informative prior distribution of N(0,10) was set to the parameters and the intercept. The reason for choosing a normal prior is that it is expected that the parameters can be negative or positive with equal probability, the mean 0 represents the null hypothesis of zero effect on the response variable and a variance of 10 makes the prior non-information. The model was fitted using the 'brms' package in R, essentially bring the merit of Stan using simple R syntax and through similar algorithms like Markov chain Monte Carlo (MCMC). 4 chains and 1000 interation (not including warmup) were used to fit the model with the help of 'brms'.

Although the marital status and residential population density provided lots of variables, a total of 8 explanatory variable is still far from any potential overfitting problem due to too many variables. Although non-informative prior does not help to deal with separation, it is expected that the variables in this model should not generate perfect prediction to the response variable. Thus no complete or quasi-complete separation should exist since most variables are dummy variables, and the results from the model supports this expectation for convergence.

Results:

TableI : Results

	Estimate	Std. Error	CI(95%)	RMSE
Intercept	-0.29	0.14	(-0.56, -0.01)	
Age	0.04	0.00	(0.04, 0.05)	
Income:$125,000 and more	0.36	0.18	(0.01, 0.70)	
Income:$75,000 to $99,999	-0.09	0.14	(-0.36, 0.16)	
Income:$50,000 to $74,999	-0.16	0.13	(-0.41, 0.08)	
Income:$25,000 to $49,999	-0.47	0.12	(-0.72, -0.24)	
Income:Less than $25,000	-0.50	0.12	(-0.75, -0.27)	
Pop. Dens.:Prince Edward Island	0.29	0.14	(0.02,0.57)	
Pop. Dens.:Rural and low population	0.37	0.06	(0.25,0.50)	
				0.3514

Table I shows the results of the model. The income level of $ 99,000 to $ 124,999 is the hidden dummy variable which appears when all five income variables equal to zero. Similarly, the population density level

of Urban and large population centers appears when all two variables for pop. dens. equal to zero. As mentioned before, the estimates for income and pop. dens. variable parameters are interpreted differently from the parameter for age.

Figure 1, Trace plot for the paramteres

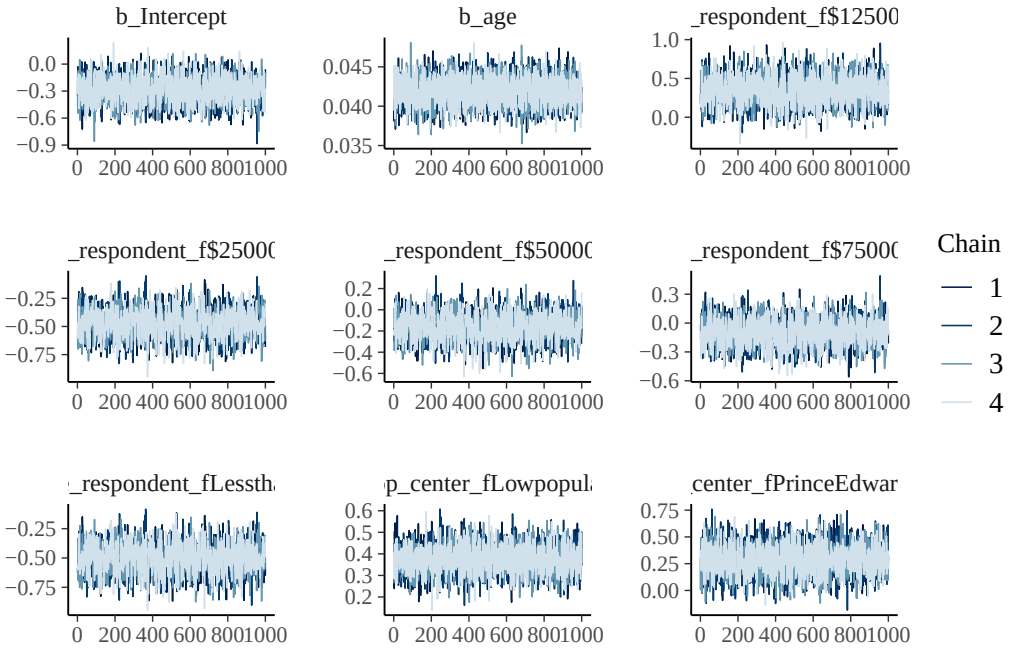


Figure 1 shows trace plots in consistent and rapid up-and-down shape with no long term trend. Meaning convergence in distribution happened rapidly and the up-and-down variation shows that the sample values are unrelated to previous ones.

Figure 2, 95% Credible Intervals for parameters

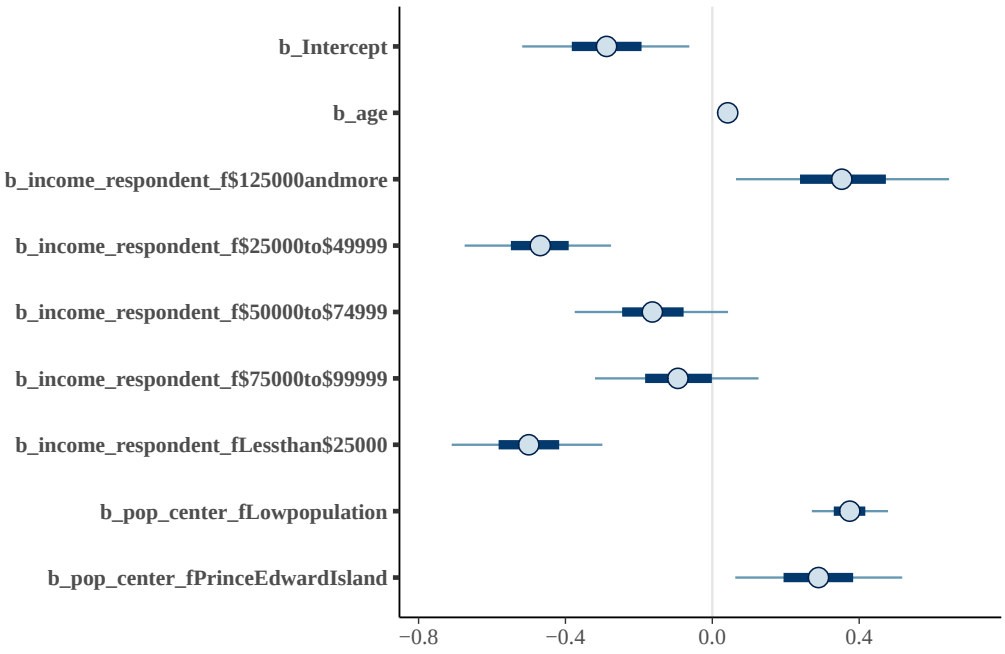


Figure 2 shows that most estimation of parameters shows a credible interval not including 0, thus rejects the null hypothesis of 0. However, the credible intervals for the income level of $50,000 to $74,999 and $75,000 to

$99,999 do contain 0, thus it failed to reject the null hypothesis.

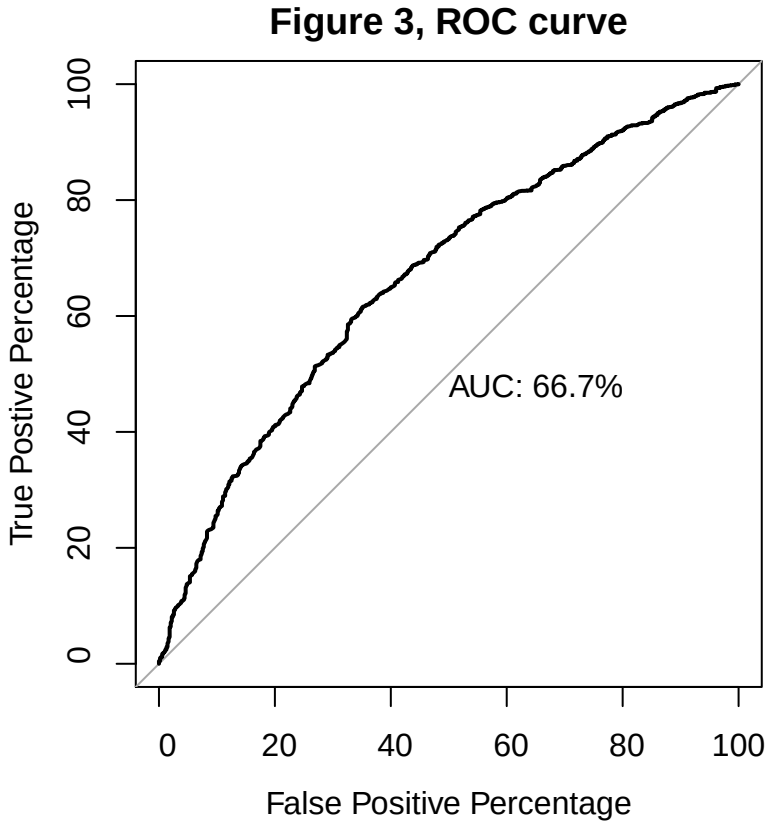


Figure 3 shows the ROC curve of the model when predicting the test dataset, with the Area Under the Curve (AUC) of 66.8%. True positive rate is the rate of successfully predicting the response variable, and false positive rate is the opposite rate. The ROC curve plots the two rates for every possible threshold of identifying the response. For example, if the threshold is 0.5, then all prediction from the model with value more than 0.5 will be classified as 1.

Discussion

Appendices

- Visualization of the test dataset of interest:

Figure A1: percentage of income level among respondents

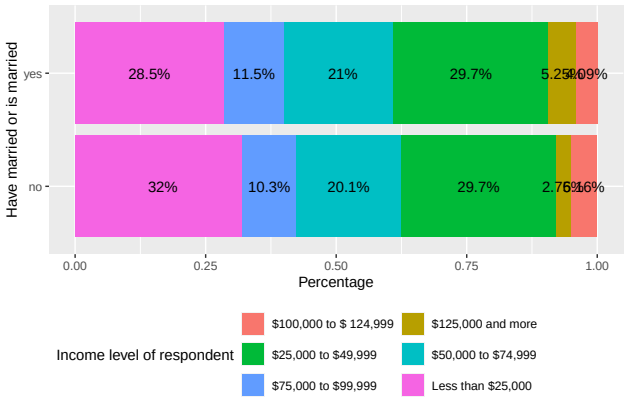


Figure A2: percentage of residential population density among respondents

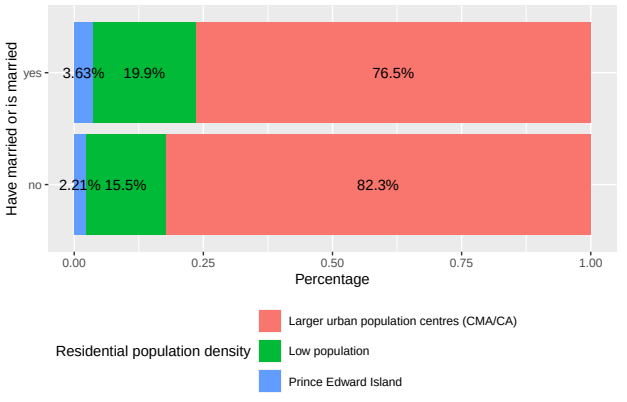
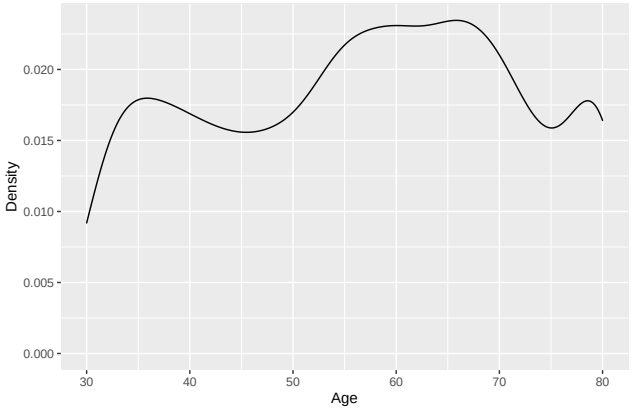


Figure A3: Distribution of age in test dataset



- Codes of the results: https://github.com/YiSu2000/GSS-Report/blob/master/GSS-Report-Su_Yi.Rmd
- Link to Statistics Canada Website on 2017 cycle of family themed GSS: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816#a4
- Survey questions on marital status, age, income level, and residential population density: https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=assembleInstr&lang=en&Item_Id=335815