Analyzing Marital Status using a Bayesian Logistic Regression Model with the 2017 GSS Dataset

Yi Su

10/15/2020

Abstract

This report focuses on the impact of income level, age, and residence location on marital status. Although it is expected that marital status should not be perfectly causal with these variables, it could be shown that the logit probability of having or had a conjugal relationship is correlated with them to some extent. A Bayesian binary logistic regression model was used to provide the results, 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. Even though the resultant model may not be very strong, the finding suggests that higher income levels do come with a higher odd of ever married at least once before, and so does age. Whereas living in the more populated area does the opposite.

1. 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 correlation between marital status, 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 socio-demographic information like age, education, and citizen status. Section 2 will discuss the framework of GSS survey in detail.

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. This fitted model will be discussed in detail in later sections.

Does this model 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. Overall, the results are beneficial for future improvements on predicting marital status. All the detailed coding and data related to these results can be found at: https://github.com/YiSu2000/GSS-Report/blob/master/GSS-Report-Su_Yi.Rmd.

2. Framework of GSS

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, which samples from different stratum randomly based on location.

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 and framework of the GSS survey can be found at the Statistics Canada website which is linked in the appendix.

Reliability of the recorded values is a strength 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. Also, the levels of categorical variables may not be great depend on the type of research. For example, the income levels would be more beneficial to this report if all incomes were measured numerically. However, that might be too private for the respondent and may not be realistic.

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, for 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 large population centers, small/low population center, and Prince Edward Island (PEI). The questions were designed by Statistics Canada. In general, the questions were carefully designed 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, reflecting a considerable amount of questions which might be overwhelming for respondents, and thus endangering the response rate and correctness of responses.

3. Data

Data cleaning of the raw dataset from Statistics Canada was done thanks to the data cleaning code created by Rohan Alexander and Sam Caetano which is referenced at the end. To choose only the observations we are interested in, all observations under the age of 30 and above 70 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, whereas any respondent with age above 70 is likely to be retired. 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 continuous variable, and the mean age of the filtered dataset is 56. In terms of other variables, annual income level is a categorical variable with six levels, with the lowest level cut at below $25,000 and highest level capped at more than $12,500. 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 got 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 training set. The training 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.

$Table I : Examples$

	Age	Marital Status	Income level(Annual)	Population density level	Ever Married	
12993	58.6	Separated	Less than $25,000	Larger urban population centres (CMA/CA)	1	yes
7664	59.7	Married	$25,000 to $49,999	Larger urban population centres (CMA/CA)	1	yes
3463	64.8	Married	$50,000 to $74,999	Low population	1	yes

Table I displays three rows of the training dataset to give an idea about the dataset.

$Table II$: Summary Statistics

Age	Ever married		Income level (Annual)	Population density level
Min. :30.00	Min. :0.0000	no :2020	$100,000 to $ 124,999: 615	Larger urban population centres (CMA/CA):8679
1st Qu.:40.80	1st Qu.:1.0000	yes:9219	$125,000 and more : 625	Low population :2169
Median :52.80	Median :1.0000		$25,000 to $49,999 :3239	Prince Edward Island : 391
Mean :51.26	Mean :0.8203		$50,000 to $74,999 :2430	
3rd Qu.:61.30	3rd Qu.:1.0000		$75,000 to $99,999 :1434	
Max. :70.00	Max. :1.0000		Less than $25,000 :2896	

Table II shows some statistics of the training dataset. The mean age is very close to the median age, thus the distribution of age is likely to be symmetric and centered on mean. Only about 22% of the training dataset have never married, this might cause some issue in slightly higher false positive rate in predicting the test set when the test set has a similar distribution. The summary of the two categorical variables are more directly presented in figure 1 and 2.

Figure 1

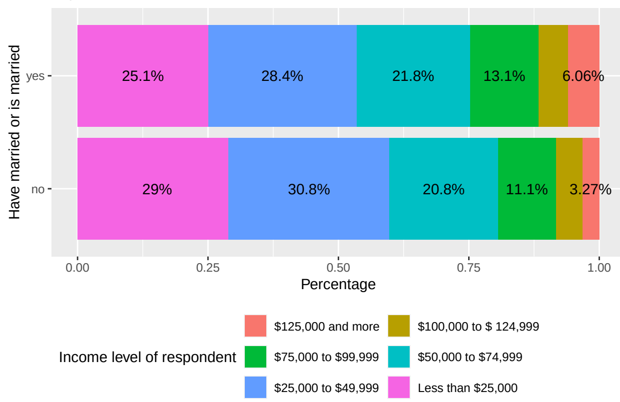

Figure 2

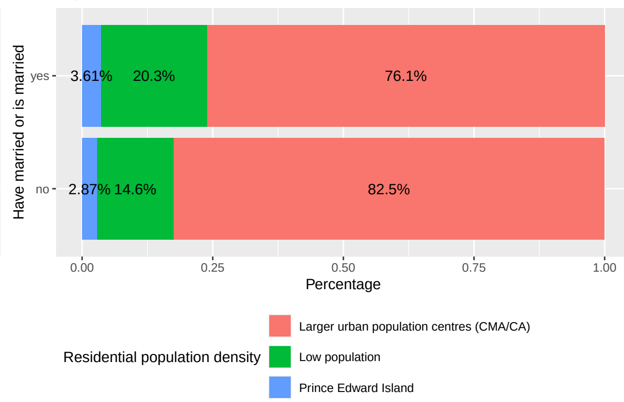

Figure 1 and 2 provides some visualization of the two categorical variables in the training dataset. Figure 1 shows that ratio of higher level income decrease progressively among samples, no matter the respondent is married or not. Around 54% to 60% of the sample has a income level lower than $50,000 in both groups. Notice that the percentage display for $100,000 to $ 124,999 level to avoid the overlap between text. Figure 2 shows that the majority of the sample lives in urban environment across both response category, and only about 17% to 24% of sample lives in smaller population regions. Overall, not much ratio wise difference arise between the married and never married sample.

4. 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. The levels of the categorical variables are followed as in the GSS dataset.

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 by modeling the logit probability as the response. Although the number of explanatory variables is essentially three, there are more variables in the model due to income level and population density level being dummy variable. There are be eight parameters for 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 how much difference is estimated between the explicit dummy variable and the implicit dummy variable holding all other parameter constant. Each of those parameter capture the fixed effect and the part of response variable explained by a specific level.

- The model may be expresses as:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 Age + \sum_{i=2}^{3} \beta_i PopulationLevel_i + \sum_{k=4}^{8} \beta_k IncomeLevel_k$$

- 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. Since the residual variance for logistic regression is the constant $\frac{\pi^2}{3}$, there's no need for a prior on it.

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). MCMC chains allow the model to develop the parameter gradually, 4 chains and 1000 iteration (not including warm up) were used to fit the model with the help of 'brms'.

Although the income levels and residential population density levels provided lots of dummy 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.

5. Results:

This section display some results from the regression model. A brief explanation to each table and figure are provided below and more discussion of the implication of these results will be included in the discussion section.

$TableIII$: *Results*

	Estimate	Std. Error	CI(95%)	RMSE
Intercept	-0.14	0.15	(-0.44, 0.16)	
Age	0.04	0.00	(0.03, 0.04)	
Income:$125,000 and more	0.47	0.17	(0.13, 0.81)	
Income:$75,000 to $99,999	0.13	0.14	(-0.15, 0.39)	
Income:$50,000 to $74,999	-0.08	0.12	(-0.32, 0.16)	
Income:$25,000 to $49,999	-0.28	0.12	(-0.52, -0.05)	
Income:Less than $25,000	-0.40	0.12	(-0.65, -0.16)	
Pop. Dens.:Prince Edward Island	0.31	0.15	(0.02,0.61)	
Pop. Dens.:Rural and low population	0.39	0.07	(0.26,0.54)	
				0.3781

Table III shows the results of the model. The income level of $ 99,000 to $ 124,999 is the default 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. Root mean squared error (RMSE) is calculated with the formula from 'brms' documentation on CRAN, using a 10 fold cross validation on the test set. The income level parameters tend to have a negative trend for any level below the default and a positive impact if above default. The population density level parameters shows a positive impact on the odds of ever got married by living in less populated region.

Figure 3

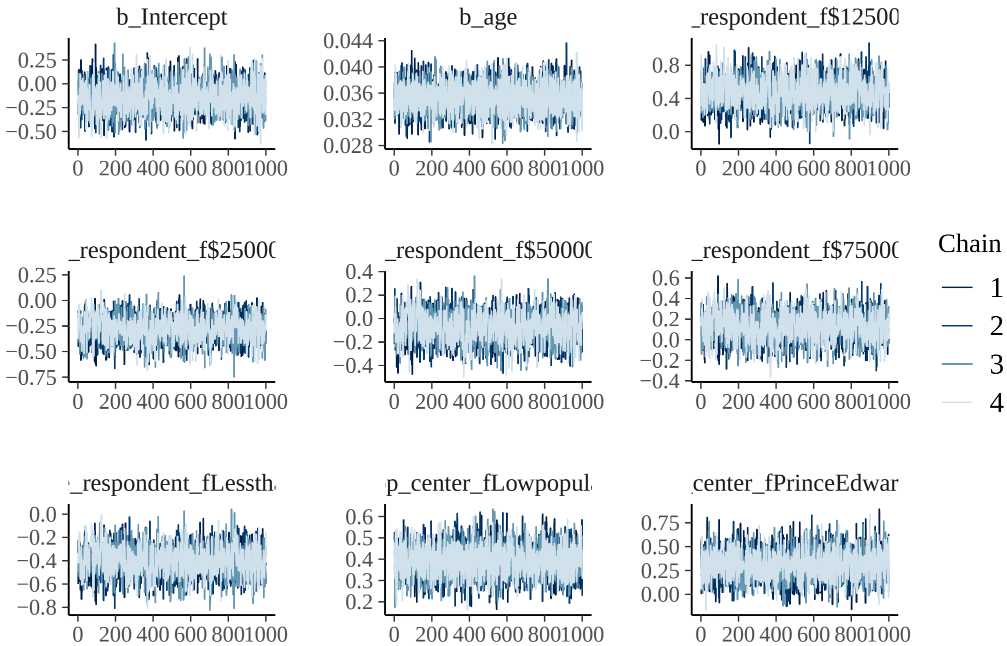

Figure 3 shows trace plots of parameters during the MCMC chains of the regression. All trace plots are in consistent and rapid up-and-down shape with no long term trend. Meaning convergence in distribution happened rapidly and the up-and-down variation shows that the sample values are unrelated to previous ones. As expected, the trace plots shows evidence of convergence in models.

Figure 4

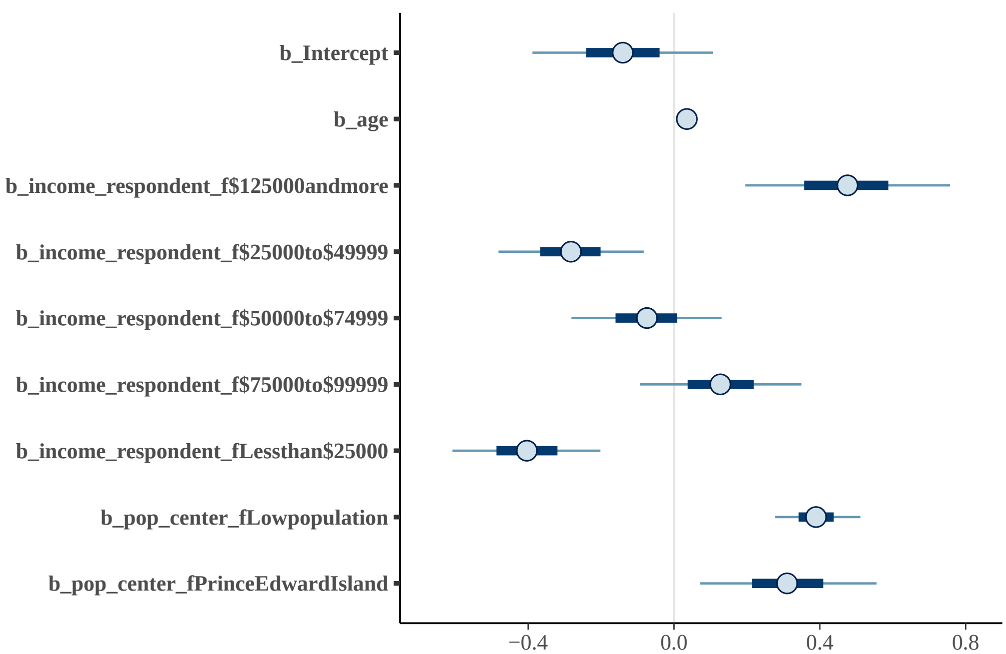

Figure 4 shows a plot of 95% credible intervals of the parameters. It can be seem that most estimation of parameters shows a credible interval not including 0, thus rejects the null hypothesis of 0. However, the credible intervals for both the intercept and the income levels of $50,000 to $74,999 and $75,000 to $99,999 do contain 0, thus it failed to reject the null hypothesis of 0 parameter value.

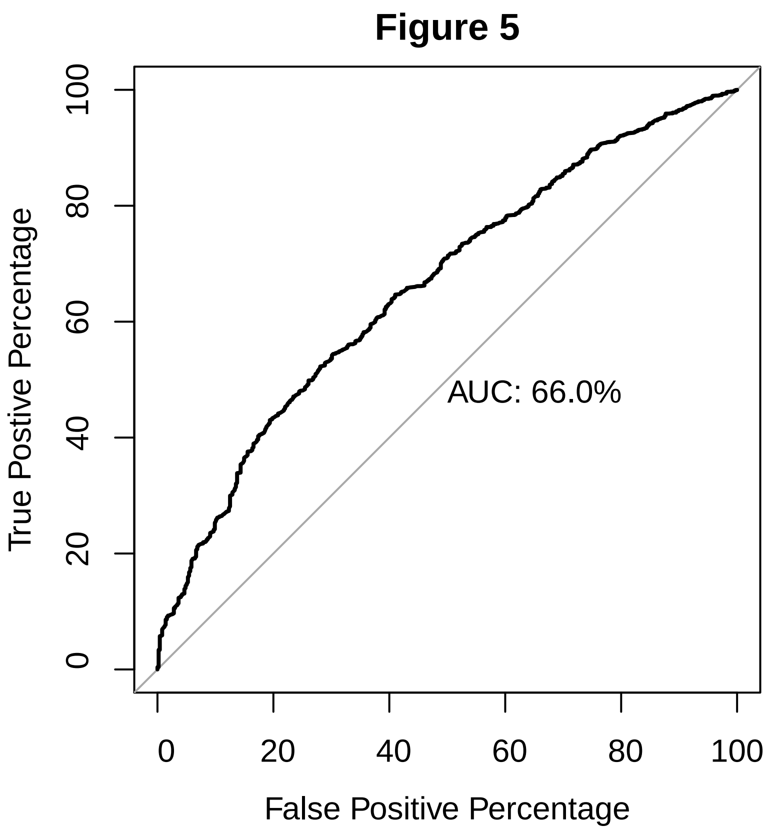

Figure 5 shows the receiver operating characteristic (ROC) curve of the model when predicting the test dataset. True positive rate is the rate of successfully predicting the response variable, and false positive rate is the opposite rate. The ROC curve plots the two rates for every possible threshold 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.

6. Discussion and Conclusion

The model suggests that income level is correlated with the probability of marrying. However, it only shows a considerable positive or negative influence when the income level is at the top or bottom of the population. This result partly follows from *Watson, T., & McLanahan, S. (2011)*, which is about relative income and marriage. They suggest that relative income is a determinant of marriage for men below a threshold income and not to men above it. The threshold income is calculated base on local groups. The resultant model also identified a negative correlation between the odd of ever married and residential population density. Meanwhile, age shows a positive impact as expected, older respondents have a higher chance of ever gotten married naturally. Besides these results, a noticeable amount of limitations exist in the model.

The resultant model may not be ideal and is not a very strong model as indicated by RMSE of 0.38 and the area under the ROC curve of 66% is generally considered poor or moderate. The RMSE of 0.38 is not considered as large since it takes the same unit as the response variable, but it suggests that the model could and should use some improvement. The ROC curve suggests that the optimum choice of true positive rate is around 65% that comes with a false positive rate of around 35% which is not low at all. Even though these indicators of quality are not in support of the model, the result of the model still provides some solid understanding of what possibly affects marital status. It can be used as an analytic model rather than a prediction model.

The intercept parameter is one of the parameters that did not show evidence rejecting the null hypothesis of zero intercepts. This should not affect our goal since the estimated intercept is a negative value and does not provide much useful information. However, the default minimum age of 30 in the training dataset might be the reason for estimating a negative intercept since the model needs to compensate for the considerable default positive impact of minimum age in the intercept.

Income level shows both positive and negative impacts on the odds of ever-married as mentioned under Table II. However, it is clear in Figure 4 that the odds of ever married is not as sensitive as expected in response to change in annual income levels. The credible intervals (at the level of 5%) of income levels below $49,999 are mostly overlapped, and this is similar for the income levels between $50,000 to $99,999. This is also the same case for the levels of population density. Thus, a reduction in levels of annual income should be appropriate when doing future improvements of the model. For levels of population density, the levels could use some rethinking because it is not a very precise leveling, a numeric version of population density in each city would likely to be the appropriate measure of this variable. However, that involves adding new data into the GSS dataset which is hard since it's not likely to get in touch with the exact same respondents.

The residential population density parameter shows an interesting result, the model indicates that the odds of ever married is higher when not living in a large population region which is a bit counterintuitive. This could be reasoned as a less populated region have a slower life pace in general, which might be in favor of developing a stable personal relationship. Meanwhile, the urban region with a large population may have the opposite effect. Also, notice that the PEI level may lack some variation in observations since there are only a handful of samples from the PEI comparing to the other two levels, a higher standard error is reflected comparing to the other parameters.

In terms of age, it shows a small but consistent and invariant positive impact on the odds of ever married. A possible explanation could be that older respondent often has more life experiences, which includes personal relationship experiences. The magnitude of age parameter might look small on its own, but it is a bit larger than expected even with age in our test dataset are ranged from 30 to 70. In compensation, most income levels below $99,999 are all predicted to have a moderate amount of negative impact on odds of ever married, and this impact becomes considerable when below $25,000. As discussed above, age is also a cause for the negative intercept estimation.

For further improvement of the model, we can consider other possible predictors such as height, education, citizen status, and even self-rated residence tidiness. The choice of a new explanatory variable is open as long as it's not in a strong linear relationship with the other variables, not a collider of the other variables, and is reasonable. However, if we are to include any variable, not in the GSS dataset, or switch to a new dataset, then there is the risk which new dataset may not be as reliable as the GSS dataset and exhibits some contradicting pattern. The modeling methodology could also be revised, especially on the prior distribution since the use of a global normal non-informative prior on parameters is not very helpful. Some informative priors could be used on different parameters based on the results from the current model like an adjusted t distribution prior to a similar but more informative distribution.

The GSS dataset is reliable in terms of a survey design as discussed in section 2, but there are of course limitations to the data. One of the limitations of observations recorded in the GSS dataset is the categorical levels as discussed above. A more accurate prediction might be made by using the numeric version of income and residential population density and even using the logged variables. The underlying bias of the present variables also exist. As mentioned in the data section, age is possibly related to income, and that was why the bottom age was raised to 30 in the modeling dataset. On the other hand, it might be the case that people living in largely populated areas are more likely to find a decent income job.

As mentioned in the introduction, even with all the good predictors and improved model methodology, marriage prediction models will never be identical to the "true" model since the "residual of emotion" will always exist. The model developed in this report is limited to the target and survey population of the GSS survey, and might not be capable to be used on any other groups without further verification. To expand this onto the board population of Canada, future developments might take advantage of the GSS survey weighting onto the whole population of Canada. Nonetheless, finding the "true" model is not the expectation for most regression models. The direction of future improvement should be providing more informative results while making more precise predictions.

Appendices

- Codes of the results: https://github.com/YiSu2000/GSS-Report/blob/master/GSS-Report-Su_Yi.Rmd
- Link to Statistics Canada Website on the 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

References

- Rohan Alexander & Sam Caetano, gss_cleaning, 7 October 2020, R codes, License: MIT, Contact: rohan.alexander@utoronto.ca
- Statistics Canada. (2017). General social survey (GSS), 2017: Cycle 31, family. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816, (Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm under DLI license through UofT)
- Watson, T., & McLanahan, S. (2011). Marriage Meets the Joneses: Relative Income, Identity, and Marital Status. The Journal of human resources, 46(3), 482–517. https://doi.org/10.3368/jhr.46.3.482

R packages:

- Ben Bolker and David Robinson (2020). broom.mixed: Tidying Methods for Mixed Models. R package version 0.2.6. https://CRAN.R-project.org/package=broom.mixed
- Gabry J, Mahr T (2020). "bayesplot: Plotting for Bayesian Models." R package version 1.7.2, <URL: https://mc-stan.org/bayesplot>.
- Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Hadley Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Jarek Tuszynski (2020). caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.0. https://CRAN.R-project.org/package=caTools
- Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret
- Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2020). naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.6.0. https://CRAN.R-project.org/package=naniar
- Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01

  Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.
- Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

- Vehtari A, Gelman A, Gabry J (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, *27*, 1413-1432. doi: 10.1007/s11222-016-9696-4 (URL: https://doi.org/10.1007/s11222-016-9696-4).
- Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.
  - 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