People's Marital Status Analysis

Haoming Hu, Ziyu Hao, QianWen Shen, SongYuan Luo

19/10/2020

Abstract

Nowadays, people are more concerned about their marriage. People care about their marriage and the factors that are associated with an intimate relationship. In this paper, we extract data from the General Social Survey to analyze what will eventually affect a marital status and cause people to divorce. Our study found out that those who have no kids and have less than 25000 yearly income are more likely to divorce, while those who are older and have no children tend to keep their marriages.

Introduction

In this paper, we gathered data about each family from the General Social Survey dataset. There are 20602 observations contained in this GSS dataset, and the variables of interest are 'province,' 'pop_center,' 'marital_status,' 'education,' 'own_rent,' 'living_arrangement,', 'average_hours_worked,' 'self_rated_health,' 'self_rated_mental_health,' 'income_respondent,' 'children_in_household,' 'ever_married,' 'religion_participation,' and 'current marriage is first'. Among all the observations, we only tend to include eligible and valid data. Therefore, we filter all the non-responses and select the groups of people who have been married, and we have 14154 observations left.

We used the generalized linear model to analyze what factors will affect people's marital status. A variable called 'divroce_indicator' was coined to show if someone has divorced, with 1 representing people having divorced and 0 representing they are still in a good relationship. Several plots and tables have been created, and they are appropriately labelled (figure 1 to 6). Among them, plot 6 is cross-validation that shows our model gives good predictions, and our non-responses are appropriately handled.

Generally speaking, this report focuses on which kinds of people are more likely to divorce. Our findings have shown that whether people have children affects the divorce rate, as people have no children are more likely to divorce. Besides, personal income plays an important role in people's marital status since low-income groups, such as those who have an income lower than 25000, have a better chance to divorce. Besides, younger people are likely to divorce compared to their elder counterparts. However, these conclusions need more to be enforced by more data because the data sampling we covered is not large enough. Therefore, further research and analysis should be done.

Data

We adopt a stratified simple random sample without replacement to make the survey a more accurate representation of the population based on each stratum. Also, it will be more sufficient. There are 27 strata in total. We are targeting those people who are 15 years of age and older in Canada, except people who live in Yukon, Northwest Territories, and Nunavut or people who live in institutions. There are 20,602 respondents chosen to take the survey. To reach those respondents, we get The Address Register and lists of

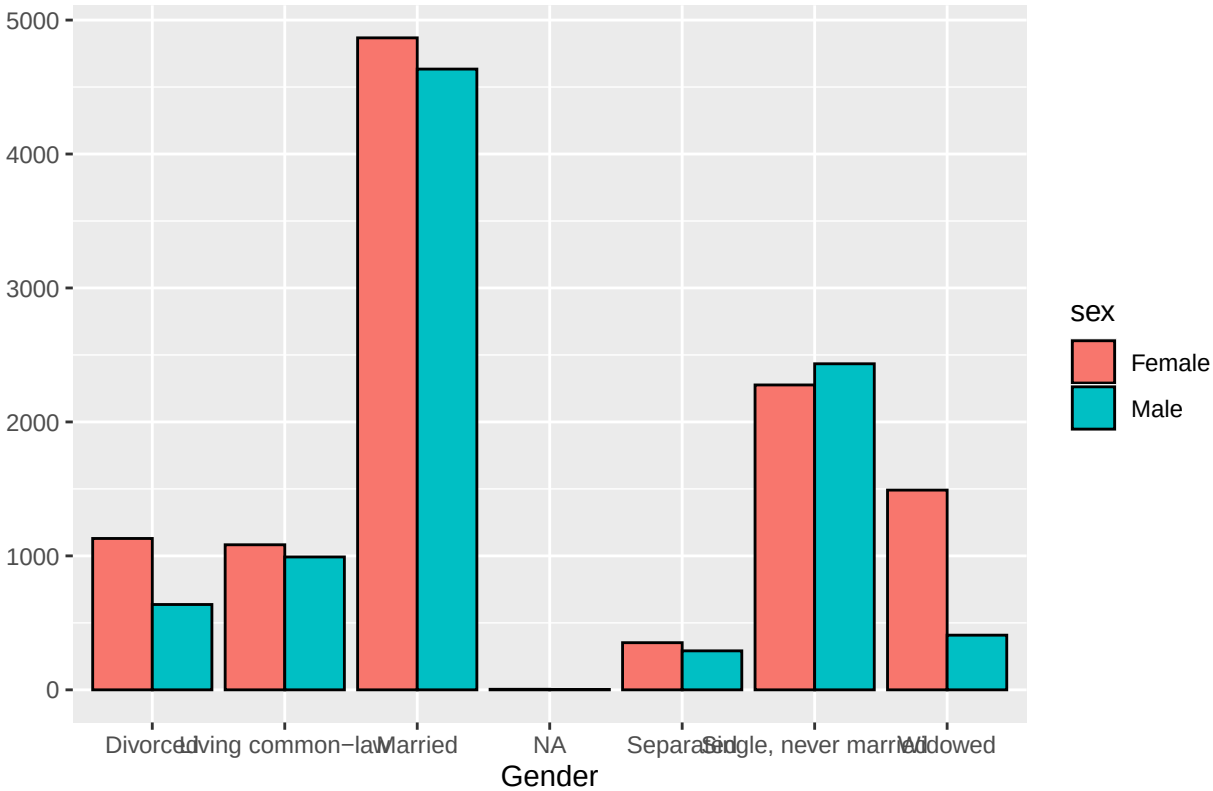
telephone numbers in use from various sources, and we use the telephone to interview those respondents. The sampling method is positive. Since it will reduce the cost and provide greater precision. In the questionnaire, we changed some questions compare to the previous survey. Some questions are deleted. For example, the income question is no longer on the survey, and we can obtain income information from the tax questions. Overall the questionnaire is highly related to our topic.

The data in use is from the General Social Survey (GSS) 2017. In this study, we aim to find out factors that influence one's marital status, more specifically, if these factors indicate whether people have divorced or not. Therefore, we created our own data set, which only contains data relating to one marital status. We include 16 predictor variables (These are values used to predict another value. In our case, we use these data to predict if people has divorced or not): 'province,' 'pop_center,' 'marital_status,' 'education,' 'own_rent,' 'living_arrangement,' 'average_hours_worked,' 'self_rated_health,' 'self_rated_mental_health,' 'income_respondent,' 'children_in_household,' 'ever_married,' 'number_marriages,' 'religion_participation," children in the household' and 'current marriage is first'.

Among these predictor variables, we filter out all non-responses and exclude people who are not married.As for our response variable (It depends on the other 14 variables mentioned above) marital status, it is divided into two groups. We denote 'Divorced' and 'Separated' as 1, denote 'Married,' 'Living common-law', and 'Widowed' as 0. In this paper, we use the mentioned 14 variables to figure out if someone has divorced, on a scalar of 0 to 1, as 1 means he/she is divorced while 0 means he/she is not.

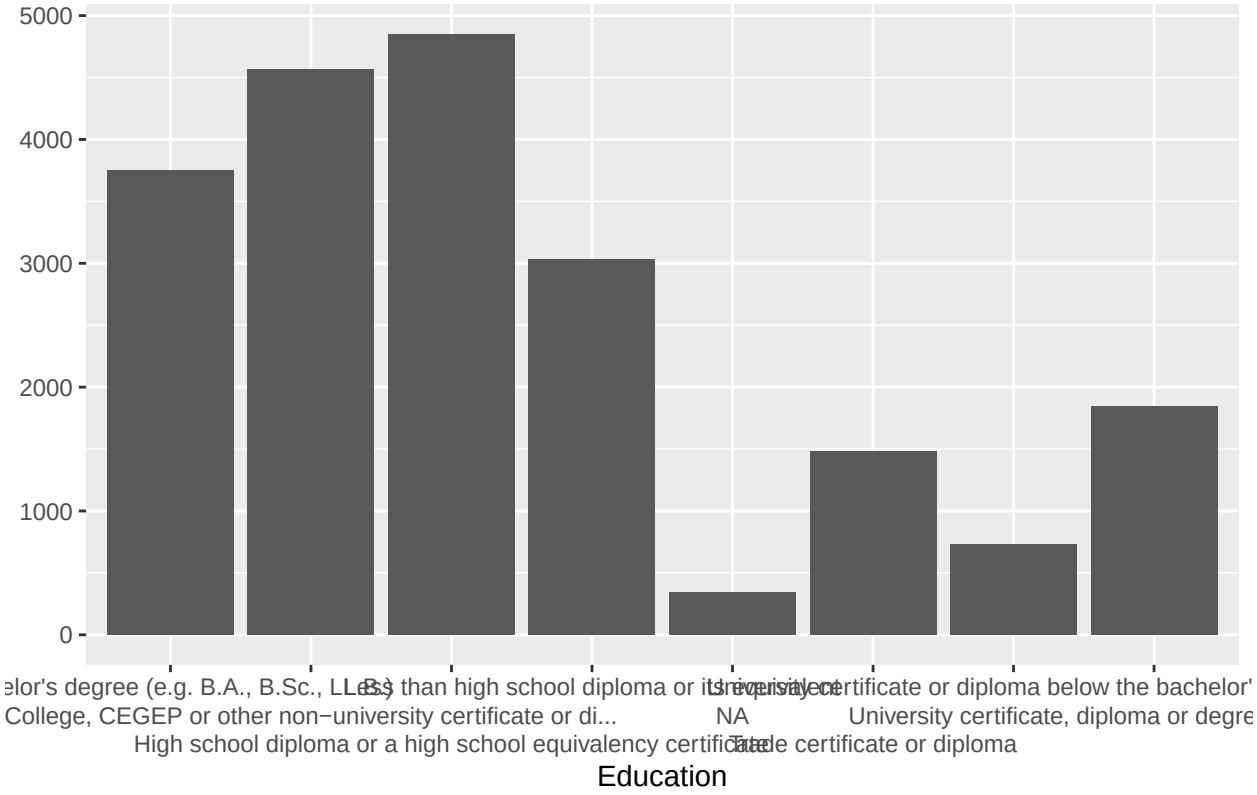
We also exclude some general data, such as age groups and gender, because they are generally normally distributed, by which I mean they will not affect our data of interest.

Figure1: Marrial Status of respondents by Gender in gss 2017



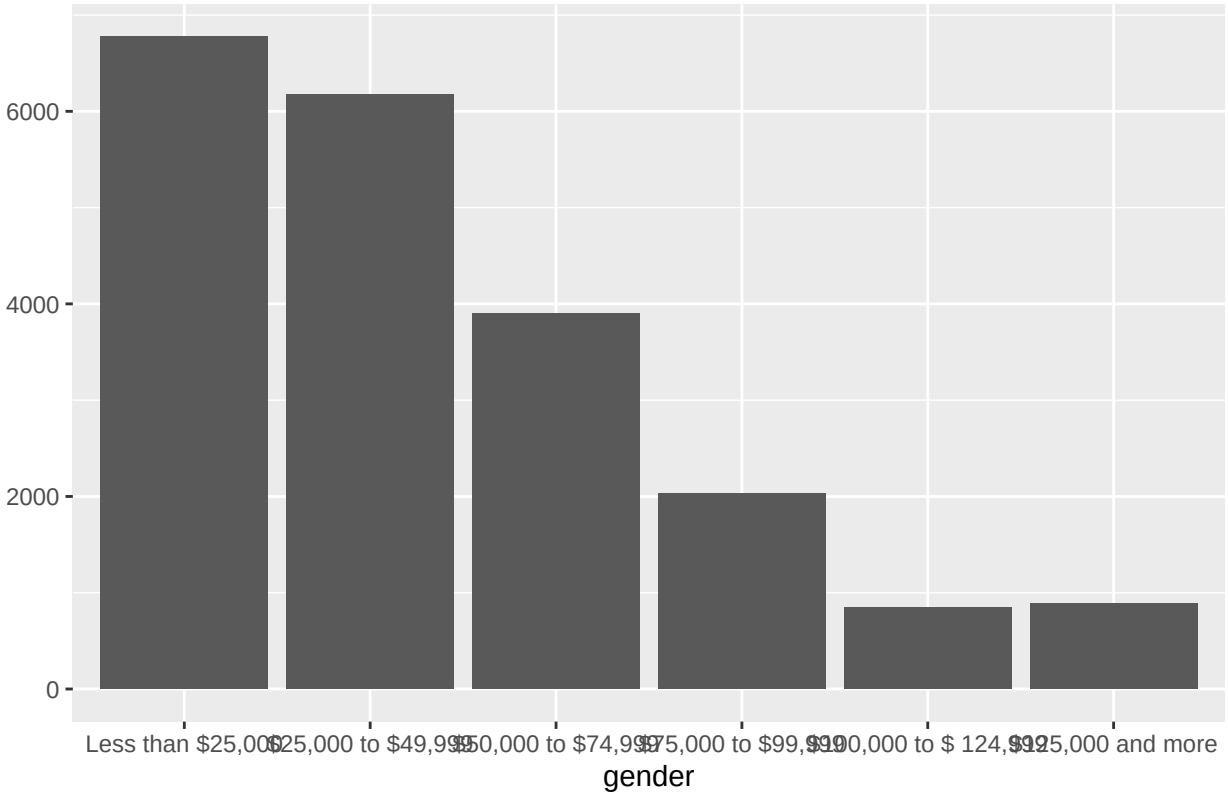
By the above graph, the number of the marital status of each gender is almost equivalent except for the widowed. Thus we could ignore gender's effect

Figure2: Education of respondents in gss 2017



As the above graph is shown, there is some missing value which will affect our research negatively. Thus we need to filter out those missing values to make our results more accurate.

Figure3: Income of respondents in gss 2017



Most respondents' income is lower than $50,000.This would give us an interesting point related to their feelings of life.

In general, the data set is good. Since it contains a large amount number of variables so that we could find factors that will affect people's marital status. However, there are many miss data in the data set. Also, the data is only collected from ten provinces, so the dataset could be biased.

Model

The model we used in this research is generalized linear model(GLM). And for the purpose of this data, we have categorized our response outcome into 2 categories, divorced or not divorced, where:

- 1 represents the number of people that are either divorced or separated
- 0 represents the number of people that are married, living common-law or widowed.

Predictors are mostly categorical, with a few exceptions, age, feelings_life, and number of marriages which are numeric variables. To be more specific. The response outcome in the dataset is divorce status, and the predictors include age, sex, income, feelings of life, education, whether the respondent owns rent, living arrangement, average hours worked per week, self rated health and mental health, number of children in the household, religion participation, whether current marriage is the first.

Our model equation is :

$$Pr(divorced = 1) = \text{logit}^{-1}\left(\alpha_{a[i]}^{age} + \alpha_{e[i]}^{educ} + \alpha_{s[i]}^{income} + \alpha_{d[i]}^{lifesatisfy} + \alpha_{e[i]}^{rent} + \alpha_{s[i]}^{living} + \alpha_{d[i]}^{workhour} + \alpha_{e[i]}^{health} + \alpha_{s[i]}^{mentalhealth} + \alpha\right.$$

After using the Logistic regression in R to investigate the relationship between martial status and the

predictors above that are possibly linked with marriage. Age, sex, two categories of education, six categories of living arrangement, 3 categories of respondent's income, 5 categories of number of children in the household, and whether current marriage is the first are the significant predictors according to their p-values.

Discussion and Results

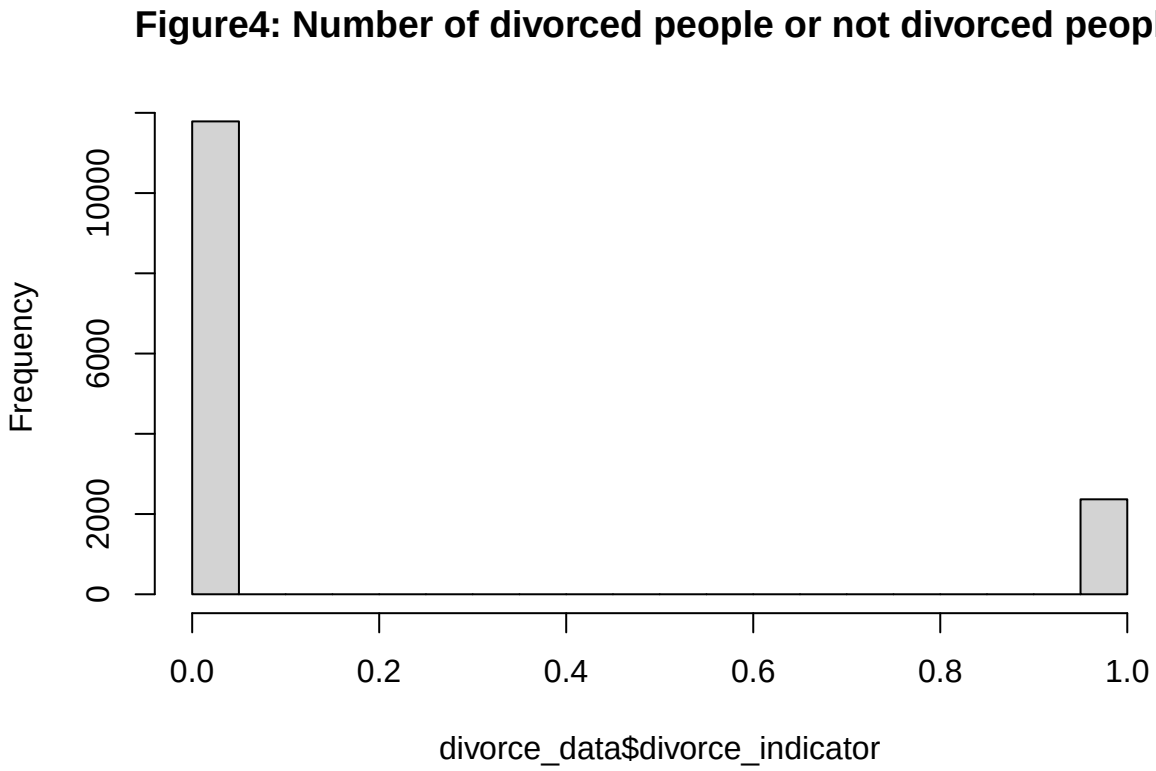
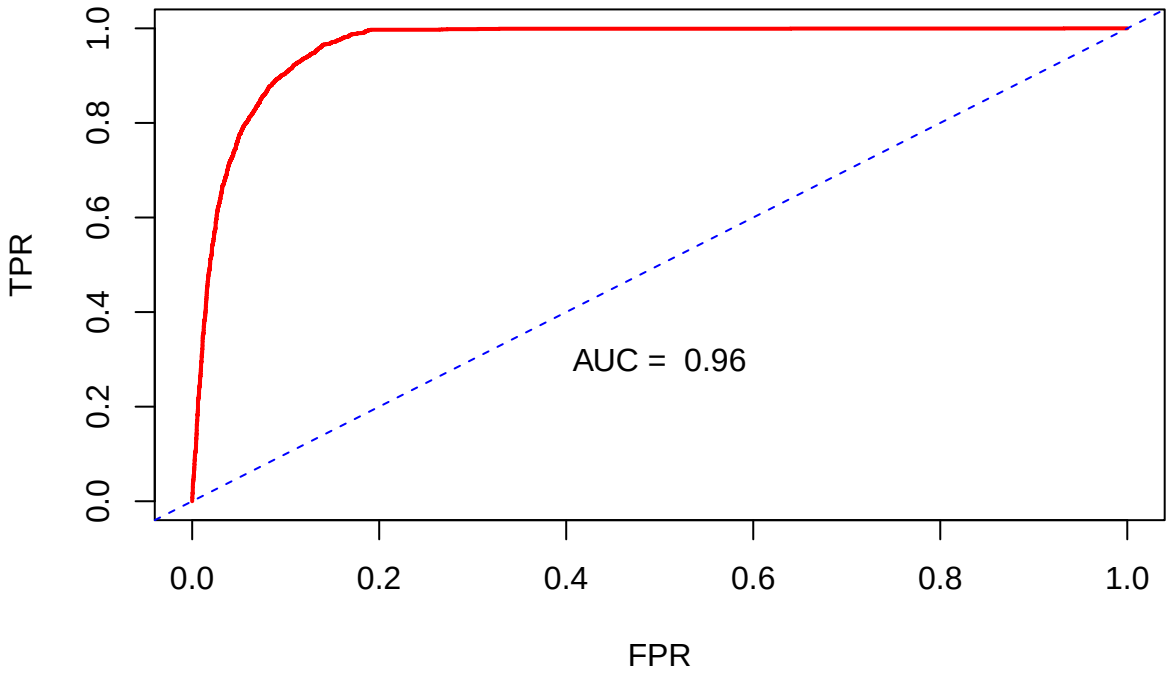


Figure 1 histogram shows that our data is significantly unbalanced, it does not follow normal distribution. Therefore, a logistic model with binomial family would be more appropriate for this data.

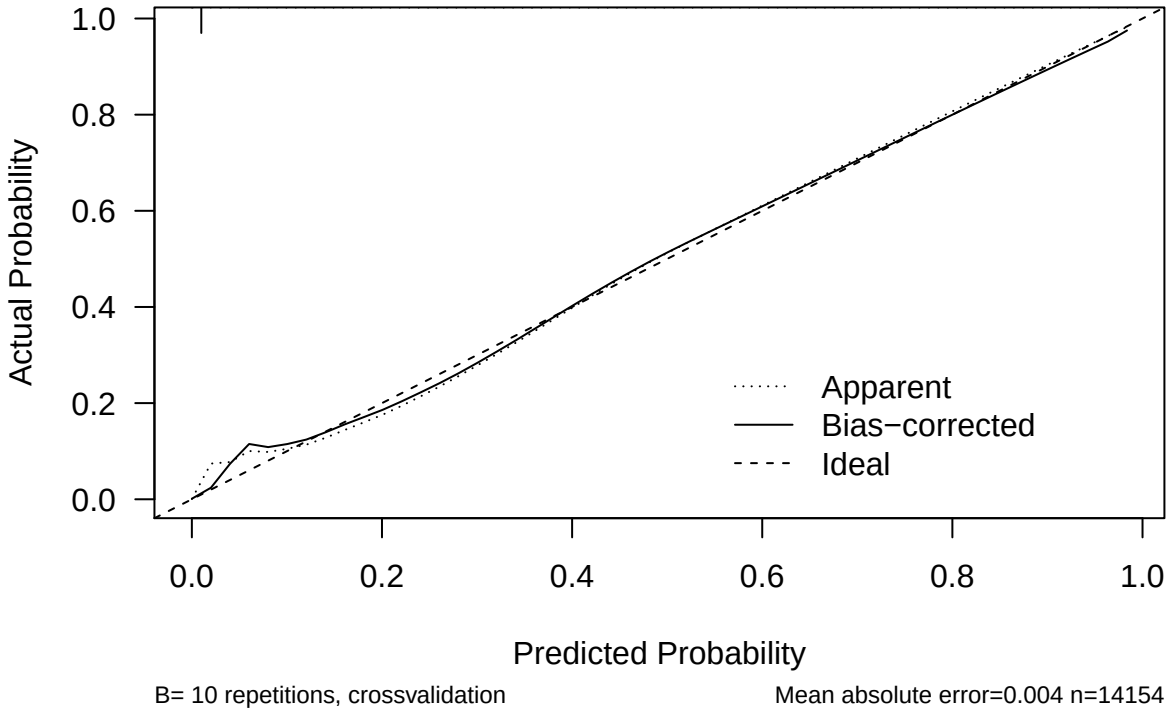
Figure 5: ROC curve



From Figure 2, the area under the curve indicates that the model predicts accurately about 96% of the time. Cross validation method is implemented to ensure our model fits the data well. See figure 3.

```
##
## n=14154   Mean absolute error=0.004   Mean squared error=4e-05
## 0.9 Quantile of absolute error=0.012
```

Figure 5: Cross Validation



As we can see from figure 3, the data are mostly consistent on the ideal line.

Weakness

An obvious drawback regarding the method of survey is that telephone interview limits coverage in certain areas, in our case, three territories namely Northwest territory, Yukon and Nunavut are excluded. Besides that, phone interviews have greatly limited not only the length but the complexity of survey questions due to respondents' lack of willingness to be involved in long telephone surveys. Lastly, an absence of visual or nonverbal cue due to phone interviews can deter disclosure of sensitive information such as "how many times have you been married?".

The method we used in coping with missing data in the variables we are interested in investigating is deletion. However, this can lead to biased estimates and therefore affecting the conclusions we might draw from the results. In categorizing the data, we included the choice "widowed" into 0, namely the non-divorced group, which affects the accuracy of our model in predicting factors affecting marriage.

The sampling method we used is Stratified sampling and the groups are 10 provinces. Instead of the generalized linear model, svyglm might be a better choice since it computes the standard errors with the consideration of the loss of precision due to sampling weights. In glm, the weight given to the errors are adjusted in the least square estimation with standard errors not being corrected.

Results

Among these predictors, age, and living arrangement with spouse are the most significant variables. From the model summary, it seems like age has a negative correlation with divorce status, which indicates that elder respondents are less likely to be divorced, however, note that this includes the widowed respondents. Also, respondent's living arrangement with no spouse had a positive correlation with divorce status, where respondent's living with spouses had a negative correlation. This indicates that Children is also an important

factor that are associated with marital status, respondents with children are less likely to be divorced. Similar conclusion can be seen from the predictor number of children.

Appendices

1. You can find our codes in: (https://github.com/songyuan-l/People-s-Marital-Status-Analysis)

References

1.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/.

2.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/

3.Frank E Harrell Jr (2019). rms: Regression Modeling Strategies. R package version 5.1-4. https://CRAN.R-project.org/package=rms

4.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

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

6.Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

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

8.dataset: https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

9.data cleaning: Authors: Rohan Alexander and Sam Caetano Contact: rohan.alexander@utoronto.ca Date: 7 October 2020 License: MIT