CE 264 Project Progress Report: Data Description and Model sketch

Chenglong Li (3032129387)
Franklin Zhao (3033030808)
Haoran Su (24375124)
Junzhe Shi (3033030938)
Kun Qian (3033030782)
Ruxin Yan (3033106689)

13 Mar. 2018

1 Introduction

As discussed in the project proposal, we are going to analyze whether couples would have a “good” marriage given some factors. Because marriage has many different status and would be affected by many objective and subjective factors, we need to declare our purpose clearly and find the factors that make most sense. The first and most important step is to identify those factors (i.e., select the variables which would significantly affect people’s marriage). Naively we could directly select those variables by hand based on our sense and some relevant research. However, such a strategy is highly unreliable and it is possible that some variables are interestingly “hidden” while they might have great impact though seems irrelevant at the first glance. Hence, we are going to select those variables (i.e., choose the best model) based on two steps: first is to use genetic algorithm (GA) roughly choosing strong variables; then based on theoretical and empirical studies around determinants and correlates of marriage status, we are going to use hypothesis testing and logit models that learned from this class to select the variables. Finally, we are going to compute the coefficients of these variables to see how would they impact the marriage of couples using logistic regression.

To begin with, we need data. Since we are analyzing the marriage, which could have tons of potential impact factors, it is really important to design and organize the survey questions so that variables can be selected from based on our data. This require large samples and a thorough research on designing those questions. Fortunately, we have already had data perfectly aligns with our study. Hence, in this report, a brief data description will be made and the design of survey will be saved, and we are going to roughly choose the factors from our data based on genetic algorithms. Finally, the sketch of our model will be made.

2 Data Description

The data used in this project is from the Japanese General Social Survey (JGSS), 2010[1]. JGSS are designed and carried out by the JGSS Research Center at Osaka University of Commerce and the Institute of Social Science at the University of Tokyo. The survey is conducted in Japanese and translated into English for the research use. The survey is conducted by two-stage stratified random sampling. JGSS s survey includes one interview and one self-administered questionnaire. Some Self-administered questionnaires require privacy considerations. The design time of each questionnaire is about 20 minutes. 9000 people take the survey from 20 to 89 years old. Because there are huge numbers of data included in the survey, JGSS used ten years to repeated cross-sectional surveys. Thus, the data may be slightly different due to the different survey years.

However, such data is sufficient for our study. The variable list is very long (rougly 500). Almost every characteristics regarding people's social information are included. It's certain that we would be able to find enough from them to study marriage. Since it is not for analyzing marriage in the first place, many variables will basically have little or nearly no impact on marriage. Hence, our first job is to select those variables based on the data. Then we are going to treat the refined data (i.e., drop other variables) it as our training data and run the logistic regression. We can first consider by ourselves to list out some important factors that may affect marriage status. And then the full variable list with comprehensive descriptions and raw data can be analyzed through GA. The complete data and data description can be found **HERE**.

3 Prior Knowledge about Marriage

According to our prior knowledge, some aspects of factors will immediately come to our mind that would affect marriage.

Firstly, we can see that economic condition will affect people's marriage. For example, getting married can sometimes help people save their living cost. However sometimes people cannot get married because they have no money. Also, many people think that the large gap of income between a couple may bring an unstable marriage.

[1] Tanioka, I., Maeda, Y., & Iwai, N. (2010), Japanese General Social Survey (JGSS). Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Secondly, social status will affect the marriage. People working in similar fields or in similar social status may share more common sense and will have a more stable marriage.

Thirdly, family background and cultural background are important. For those people from areas with open mind, they usually keep an open mind to marriage which means a higher single or divorced rate. However, for those conservative people, they tend to keep their marriage. Parents' marriage is usually a sample for children to learn about marriage. If their parents have a good marriage, then children will tend to believe in marriage. However, if parents get divorced, children tend to lose belief in marriage.

In fact, there are many other factors such as health status, education background, political view and mental health that would affect the marriage status. Here we will not cover all of them. We will analyze briefly with data and combine our analysis with common sense to present a simple sketch or our model.

4 Preliminary Selection Using GA

Genetic algorithms mimic the Darwinian natural selection process to solve optimization problems[2]. Every candidate solution corresponds to an individual of a population and is represented by its genetic code, referred to as chromosomes in the following discussion. Each individual's genetic code embodies one candidate solutions to the optimization problem. By breeding among individuals who have genetic code with better fitness in terms of the optimization problem, genetic algorithms allow the population to evolve and become increasingly fit.

"**GA**"[3] is an R package developed by group member Franklin and 3 other students last semester to implement the genetic algorithms. In this part, we are going to use *select*() in this package to roughly select the potential variables from the data. Code and results are shown in **Appendix**.

[2] Givens, G.H., & Hoeting, J.A. (2012), Computationa statistics (Vol.710). John Wiley & Sons.

[3] Zhao, F., Zhong, M., Jabbari, A., & Genies, C. (2017), An R package for variable selection in regression problems based on genetic algorithm. Statistics 243 Final Project.

5 Sketch of Model

After running GA, we noticed that the result is roughly align with what we expected. Hence, using the results, we now make a sketch of model based on theoretical and empirical studies around determinants and correlates of marriage status, especially studies on the Japanese society[4,5]:

Sex (SEXA): gender is an important indicator, because males and females usually pose different opinions about marriage (male is 1 and female is 0)

Citysize×M (municipality) (SIZE): the region that people live in has an impact on their attitude toward marriage, usually people living in mountain region and small towns are conservative but people living in big cities are open to divorce or staying single; we consider city size with 4 levels (from 1 to 4 means city size grows larger) also municipality is also assigned into several levels (3), multiply them together as an indicator about the characteristic that people live in.

Year (Birth year) (SIZE2K): people grow us in different time period will share different values; we will allocate different time periods, for example, 40s to 50s will be a level, 50s to 60s will be a level and so on.

$\mathbf{E}_{\mathbf{level}}$ (Education level) (XXLSTSCH): people with different education levels may also have different opinion about marriage, high school and lower= 1, undergraduate = 2, graduate school and higher = 3.

$\mathbf{W}_{\mathbf{hours}}$ (XJBSCH): working hours every week: working hours every week means a lot, because if people dont work, it means their special social status, if they work a lot, they will have limited time to be with their family which is not good for a health family environment.

$\mathbf{W}_{\mathbf{income}}$ (INCMAN): working income: working income means a lot, because from an economy point view, marriage is a way to ease peoples financial burden because they can share lot of things together. So income would be a very important factor.

[4] Takahashi, K., Nin, T., Akano, M., Hasuike, Y., Iijima, H., & Suzuki, K. (2017). Views of Japanese medical students on the work-life balance of female physicians. International journal of medical education, 8, 165.

[5] Sasaki, S. (2017). Empirical analysis of the effects of increasing wage inequalities on marriage behaviors in Japan. Journal of the Japanese and International Economies, 46, 27-42.

$\mathbf{Diff_{income}}$ (INCMAIN): the income difference with husband or ex, if single then dons consider; Sometimes the difference between couples income will lead to divorcement. Because the big difference in income indicates the different social status. Usually people from different social rank share few same values which will lead to divorcement.

$\mathbf{Num_{child}}$ (CCNUMTTL): number of children: usually, if people have already had children, they tend to stay or go into a marriage.

$\mathbf{iNum_{child}}$ (APPCUMNN): ideal number of children: Usually people love children or want children want a warm family which will lead them to be in a marriage.

$\mathbf{P_{view}}$ (political view) (OP5RADCA): we set neutral, aggressive and conversive as three different levels represented by 1,2,3; Political view is a very important factors. For example, usually people who support Democratic share a total different value system with people who support Republican. This will reflect on the peoples marriage view.

$\mathbf{P_{health}}$ (physical health) (OP5HLTHZ): peoples physical status will have an effect on peoples tendency to get married, suppose if a person is in bad physical status, then it will be more difficult than others for him or her find a true love; here we will set four levels (0,1,2,3) to indicate the health status.

$\mathbf{M_{health}}$ (mental health) (SFHLCND): mental health is extremely an important indicator because for those unlucky people who is depressed, they even dont want to find their love. Similarly, we will set four levels for the mental health status (0,1,2,3)

Willingness (willingness to ask others for help when in trouble) (Q4WWHPHH): this is a very interesting thing to look into. Because for many people, they tend to ask others for help. For other people, they are unwilling to ask others for help. In fact independent people tend to solve difficulties by themselves thus it would be easier for them to stay single.

According to the analysis above, a possible utility function is listed as below:

$$\begin{aligned} U = \; & \beta_{sex} \times Dummy_{sex} + \beta area \times Dummy_{citysize} + \beta_{year} \times Dummy_{birthyear} \\ & + \beta_{edu} \times Dummy_{edulevel} + \beta_{whour} \times hours_{work} + \beta_{income} \times income_{work} \\ & + \beta_{diffinc} \times Diffincome + \beta_{child} \times Num_{child} + \beta_{ichild} \times iNum_{child} \\ & + \beta_{politics} \times Dummy_{pview} + \beta_{ph} \times Dummy_{physicalhealth} \\ & + \beta_{mh} \times Dummy_{mentalhealth} + ASC_{willingness} + \varepsilon \end{aligned} \tag{1}$$

Future work will focus on refining the model using pylogit.

Appendix

GA code and results:

```
devtools:::install_github('QinganZhao/GA')
load('JGSS')
jpY ← jp[, c('MARC')]
Yjp ← as.matrix(log(jpY))
jpX ← jp[, !(names(jp) %in% c('MARC'))]
Xjp ← as.matrix(jpX)
GA:::select(Xjp, Yjp, reg='glm')
```

```
Selected predictors:
SIZE SEXA TP5UNEMP AGESTPWK SZSJBHWK TPJOBP TPJBDP
SZCMTHR XXJOB XJOBDWK SZTTLSTA ST5JOB DOMARRY SSJB1WK SSTPUNEM
SSSJBHWK SSTPJOB SSTPJBDP SSTPJBSE SSXJBSCH SSSZWKYR SSSZSTFA SPAGEX
SPLVTG PPLVTG MMLVTG PPAGE MMAGE MMJOB CCNUMTTL CC01SEX CC01AGE
CC02LVTG CC02AGE CC02MG CC03MG CC03JOB CC04LVTG CC04MG CC05SEX
CC05JOB CC06SEX CC06AGE CC07SEX CC07LVTG CC07AGE CC07JOB CC08MG
CC08JOB SZFFOTHR FFH01REL FFH03SEX FFH04REL FFH04SEX FFH05REL FFH05SEX
FFH07REL FFH07SEX FFH07AGE SZFFONLY SZFFTTL FFHEAD SZFFOUT FF001REL
FF001WHY FF002REL FF002WHY FF003REL FF003WHY FF005REL FF005WHY
FF006REL FF006WHY INCSELF INCSP INCPEN INCUEB INCIRR INCRENT INCMAIN
SZINCOMA XNUMSISE XNUMBROY XSSNBROY XSSNSISY PREF15 TP5LOC15 PPJBXX15
PPJBSZ15 MMJBTP15 XXLSTSCH SSLSTSCH PPLSTSCH DOLSTSCH XGRADE XSPSCH...
<truncated>
Fitness value: -28489.5
```