Whether Males in BC Have Married or Not Is Affected by Their Age and Feelings of Life and Self-rated health

Huayu Wu, Jingyi Nie, Qian Wang, Yueran Hu

2020/10/19

Contents

The Github link is https://github.com/Lily-WangQian/ProblemSet_3_G_78.git

We use `R`(R Core Team 2020), `dplyr` package (Wickham et al. 2020), `ggplot2` package (Wickham 2016), and `gridExtra` package (Auguie 2017) to do the analysis.

1 Abstract

In this research, we used data from Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the Family to explore whether the marriage of males who is 15 years of age and older in BC has a relationship with their ages, feelings about life and the feelings of their health. The results show that the relatively elder males with better feelings about life and self-evaluated health as "very good" and "excellent" are more likely to get married. The reason to research on this topic is that marriage may be a potential factor that affects social stability and the population ageing problem.

2 Introduction

According to a study from the scholars of German Medical Center(DePaulo 2019), they found that satisfaction with single lives increases with age. However, this phenomenon is not conducive to some developed countries

like Canada, since ageing population problem is getting serious in Canada(Clemens and Parvani 2019). With the lower marriage rate, the birth rate will decrease meanwhile, and the ageing problem will become more serious. Therefore, to know the relationship between people's ages, feelings about life, feelings of health and their marriage is important.

We used data from Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the Family to explore whether the marriage of males who is 15 years of age and older in BC has a relationship with their ages, feelings about life and the feelings of their health. The data of GSS was collected in 2017, the population of this data is all Canadian with 15 years older among 10 provinces. In this report, we focused on three factors that may affect the marriage of males - ages, feelings about life as a whole and self-rated health. After getting rid of the invalid data, we divided the data into five self-rated health levels to see their different age distributions and life satisfaction distributions. Additionally, we got the model between age and feelings about health and found that the elderly are more likely to rate themselves with relatively lower health levels. Then, we used the logistic regression model to predict the relationship between the marriage of males in British Columbia and their ages, feelings about life and feelings of health. We found that the relatively elder males with better feelings about life and self-evaluated health as "very good" and "excellent" are more likely to get married.

However, in this research, our result has some limitations and we still need to improve our research. The first weakness is that we do not consider some lurking variables such as males' financial status and if they ever been divorced. Additionally, the questions we asked in the questionnaire were too subjective, such as "feeling about life", different people have different criteria, and it is hard for us to unify them. Meanwhile, the missing values may also influence our result. Thus, we still need to improve the research to get a more accurate and realistic result.

3 Data

Our dataset is gathered from the Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the family, which is a sample survey with cross-sectional design. In order to carry out sampling, the 2017 GSS used the stratification methodology to collect and sample data based on the geographic areas(Canada 2017). To be specific, each of the ten provinces in Canada was divided into strata(Canada 2017). The population of the survey is all Canadian 15 years of age and older, excluding full-time residents of institutions and residents of the Yukon, Northwest Territories, and Nunavut(Canada 2017). Besides, the two distinct components of the survey frame are the list of all residences within the ten provinces in Canada and the list of telephone numbers which is available to Statistics Canada(Canada 2017). Since the aim of our research is to predict whether the marriage of males who is 15 years of age and older in BC is related to their ages, feelings about life and the feelings of their health, our sample is males 15 years of age and older in British Columbia. The information of respondents was collected via computer assisted telephone interviews scheduled on the daytime of a week. Significantly, people who at first rejected to take the survey would be re-contacted up to two more times to encourage them to participate by explaining the importance of the survey(Canada 2017). In order to avoid the time conflict, interviewers would make an appointment at a convenient time with participants to call back(Canada 2017). Additionally, the first key feature of the survey is that it uses the stratified sampling method to collect data, which means each stratum within its province corresponds to a record in the frame(Canada 2017). Another key feature is the survey frame is not created using single-source but using various available linked sources, such as Census of population(Canada 2017). Moreover, the strengths of the survey are the sample size of data is large (about 20602) and the survey does not have non-sampling errors(Canada 2017). However, there also exist some limitations in the survey, such as the potentially high cost and the long spent time of the telephone interview. In our dataset, we mainly used ever_married, which is the response variable, and age, feelings_life, self_rated_health, which are explanatory variables. In order to avoid the potential influence of some values, especially "NA" and "Don't know" in self_rated_health, we filtered our original data to only show the other five levels in self_rated_health. Since we focused on researching males in British Columbia, we established a new dataset by filtering the data to only show cases that province and sex are British Columbia and male respectively.

1. Description of age group by self-rated health

For the age of this dataset, we divided the data based on self-rated health into five levels to see their different age distributions. Table 1 indicates, for self-rated "Excellent" health levels, the range of age is from 15.1 to 80 years old. The mean age of excellent health level is 48.00 years old, and the median is 47.35, which means it is not a perfectly symmetrical distribution. The interquartile range is 30.76. For self-rated "Good" health levels, the range of age is from 15.5 to 80 years old. The mean age of good health level is 54.23 years old, and the median is 56.55. The interquartile range is 22.77. For self-rated "Poor" health levels, the range of age is from 22.8 to 80 years old. The mean age of poor health level is 61.29 years old, and the median is 64.55. The interquartile range is 20.70. We see that participants with younger age would rate themselves higher health levels. That means age may have a positive relationship between health levels.

Table 1: Summary of age group by self rated health

self_rated_health	min	qrt_1st	mean	median	qrt_3rd	max
Excellent	15.1	33.42	48.00	47.35	64.18	80
Fair	18.2	42.93	57.24	60.45	71.03	80
Good	15.5	41.85	54.23	56.55	67.62	80
Poor	22.8	53.82	61.29	64.55	74.55	80
Very good	15.0	34.85	50.83	51.80	67.45	80

2. Description of feelings about life as a whole group by self-rated health

For feeling about life of this dataset, we divided the data based on self-rated health into five levels with 'don't know' and 'not available' data to see their different life satisfactions distributions. Table 2 shows, for self-rated "Excellent" health levels, the range of life satisfactions is from 0 to 10. The mean feeling about life of excellent health level is 8.68, and the median is 9, which means it is not a perfectly symmetrical distribution. The interquartile range is 2. For self-rated "Good" health levels, the range of life satisfactions is from 0 to 10. The mean feeling about life of good health level is 7.90, and the median is 8. The interquartile range is 2. For self-rated "Poor" health levels, the range of life satisfactions is from 0 to 10. The mean feeling about life of poor health level is 5.66, and the median is 5. The interquartile range is 4. We see that participants with a worse feeling about life would rate themselves lower health levels. That means feeling about life may have a positive relationship between health levels.

Table 2: Summary of feelings about life as a whole group by self rated health

self_rated_health	min	qrt_1st	mean	median	qrt_3rd	max
Excellent	0	8	8.64	9	10	10
Fair	2	5	6.83	7	8	10
Good	0	7	7.90	8	9	10
Poor	0	4	5.66	5	8	10
Very good	0	7	8.25	8	9	10

4 Graphs

Figure 1 shows the distribution of age is not symmetric. We see that the prominent peak lies to the right with the tail extending to the left, that means it is a left-skewed. The variable of age is not approximately normal. Moreover, there are about 80 participants who are the age of 80, which takes the biggest quantity of all participants.

Figure 2 demonstrates the distribution of feelings about life as a whole is not symmetric. We see that the

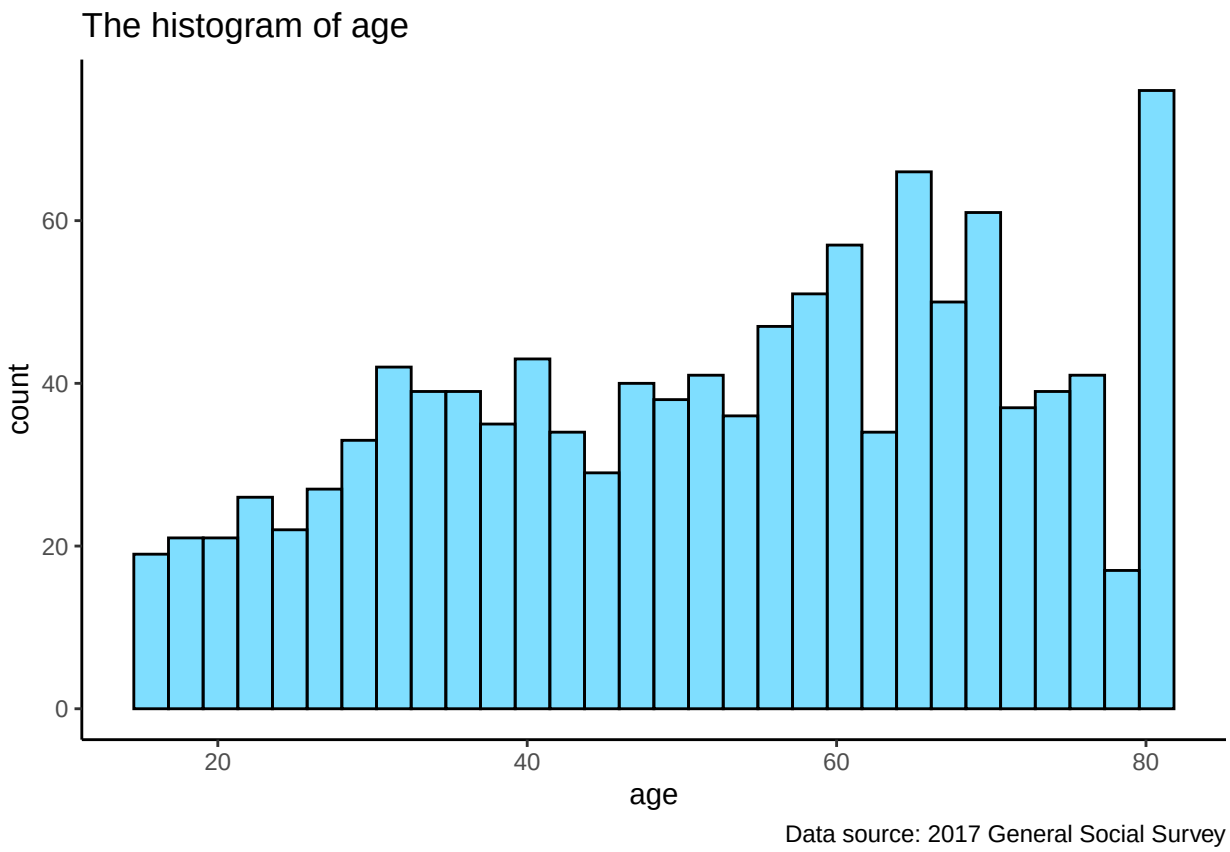


Figure 1: fig1

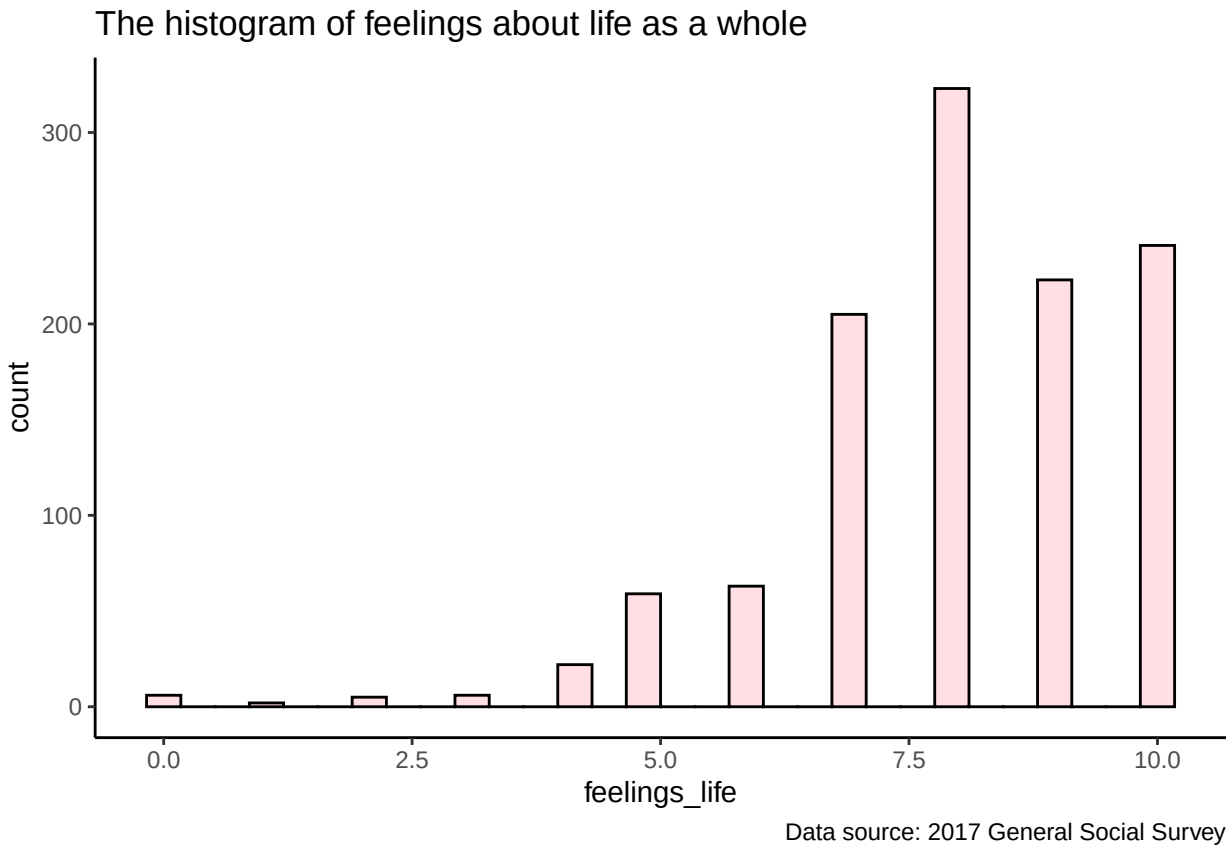


Figure 2: fig2

prominent peak roughly lies to the right with the tail extending to the left, that means it is an approximately left-skewed. The variable of feelings about life as a whole is not approximately normal. There are more than 300 participants who rated their feelings about life 8, which takes the biggest quantity of all participants.

The scatter plots of age and feelings about life as a whole group by self rated health

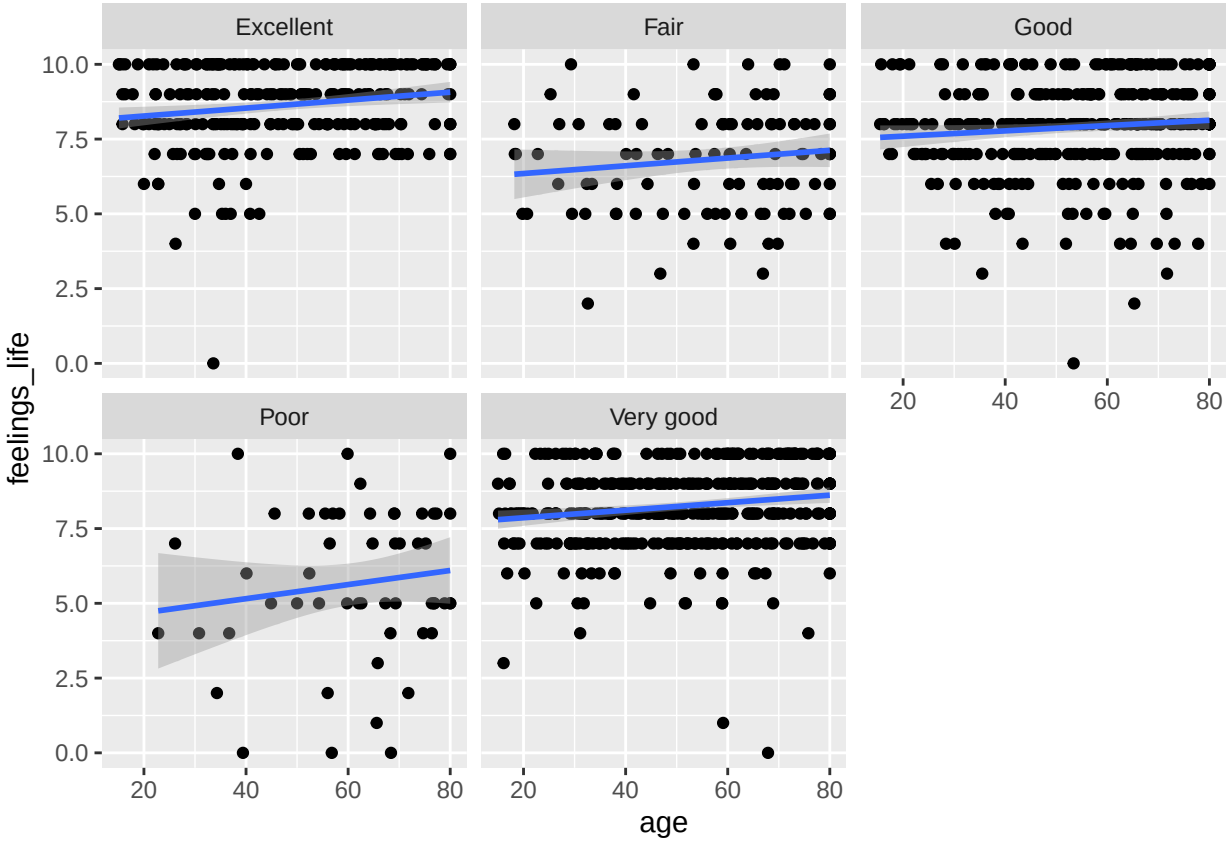
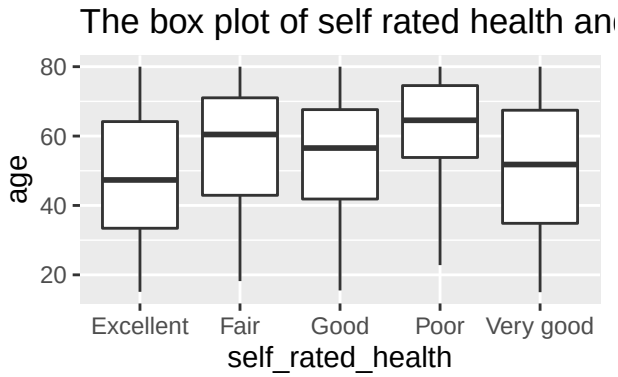
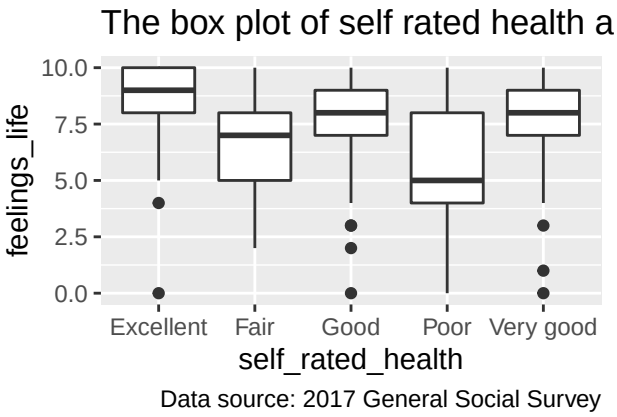
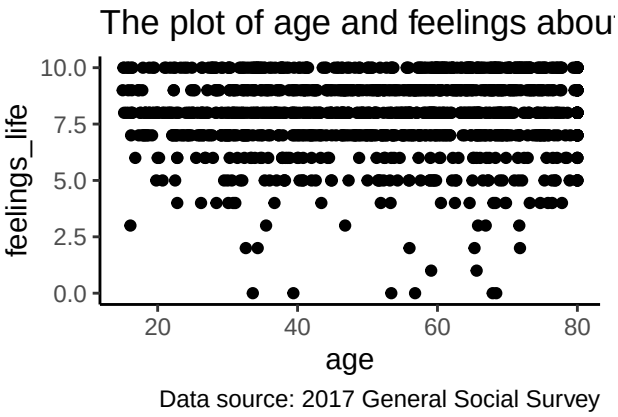


Figure 3: fig3

Firstly, by observing Figure 3, we found that the distributions of all points divided by self-rated health are not diffuse. Secondly, we observe that age and feelings about life as a whole may have a weak linear relationship between each other in every self-rated health level. Thirdly, we can see a weak positive direction form all six plots. Among six self-rated health levels, we know that "Poor" health level has the largest slope. That means participants with higher age would rate themselves worse health levels, especially for people who rated their health levels as "poor".







For the bar plot of self-rated health, we see that "Very good" self-rated health level takes the maximum population about 400, followed by "Good" self-rated health level and "Excellent" self rated health level. That shows people generally consider themselves healthy.

For the box plot of self-rated health and age, we see that participants' ages are approximately normal distribution and approximately symmetric with a slight right-skewed in five self-rated health level. There is no outlier in data. The median of all five self-rated health levels is between 45 and 65 years old. "Poor" self-rated health level has the highest median age.

From the box plot of self-rated health and feelings about life, we found that participants' feelings about life may not be a normal distribution. There are some outliers in "Excellent", "Good", and "Very good" self-rated health levels, which might affect the result of the distributions.

Lastly, the plot of age and feelings about life as a whole group shows that the distribution of all points is an even distribution. We cannot find a linear relation between each other.

** research question: predict ever married among males in BC **

```
##
## Call:
## glm(formula = ever_married ~ age + feelings_life + self_rated_health_b,
##     family = "binomial", data = new_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6498  -0.6914   0.3815   0.7127   1.8867
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
```

Table 3: Summary of model estimates

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.232	0.536	-7.901	0.000
age	0.080	0.005	15.483	0.000
feelings_life	0.133	0.048	2.750	0.006
self_rated_health_bFair	-0.240	0.463	-0.518	0.605
self_rated_health_bGood	-0.025	0.419	-0.059	0.953
self_rated_health_bVery good	0.247	0.426	0.580	0.562
self_rated_health_bExcellent	0.315	0.444	0.710	0.478

```
## (Intercept)                 -4.23157    0.53557  -7.901 2.76e-15 ***
## age                          0.08005    0.00517  15.483  < 2e-16 ***
## feelings_life                0.13327    0.04847   2.750  0.00596 **
## self_rated_health_bFair     -0.23977    0.46324  -0.518  0.60474
## self_rated_health_bGood     -0.02490    0.41878  -0.059  0.95260
## self_rated_health_bVery good  0.24694    0.42556   0.580  0.56173
## self_rated_health_bExcellent  0.31501    0.44364   0.710  0.47766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1428.4  on 1154  degrees of freedom
## Residual deviance: 1058.8  on 1148  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 1072.8
##
## Number of Fisher Scoring iterations: 5
```

The logistic regression we illustrated here describes the linear relationship about how all the explanatory variables which are respondents' age, the feelings about their life, and their self-rated health relate to the response variable which is whether the respondent ever legally married. From the positive coefficient of the age and feelings_life, it is clear that the respondent's age and their feeling about their own life increase the odds that the respondents have legally married. For the five levels of how respondents rate their health, compared to the evaluation of poor, the respondents who rate themselves have fair or good health tend to have fewer odds of having married legally in BC. By contrast, the respondents who think themselves have a very good and excellent health increase the odds of having married. Moreover, according to Table 3, we could know the estimate, standard error, test statistics, and respective p-value for each explanatory variable

5 Model

$$\begin{aligned} y_i = &- 4.232 + 0.080x_{age,i} + 0.133x_{feelings_life,i} - 0.240x_{self_rated_health_2,i} \\ &- 0.025x_{self_rated_health_3,i} + 0.247x_{self_rated_health_4,i} + 0.315x_{self_rated_health_5,i} + \epsilon_i \end{aligned}$$

$$\begin{aligned} Y_i = &- 4.232 + 0.080X_{age,i} + 0.133X_{feelings_life,i} - 0.240X_{self_rated_health_2,i} \\ &- 0.025X_{self_rated_health_3,i} + 0.247X_{self_rated_health_4,i} + 0.315X_{self_rated_health_5,i} + \epsilon_i \end{aligned}$$

$$\begin{aligned} \log \frac{p_i}{1-p_i} = &- 4.232 + 0.080x_{age,i} + 0.133x_{feelings_life,i} - 0.240x_{self_rated_health_2,i} \\ &- 0.025x_{self_rated_health_3,i} + 0.247x_{self_rated_health_4,i} + 0.315x_{self_rated_health_5,i} + \epsilon_i \end{aligned}$$

Here is the final logistic regression model we have estimated. The family of the response variable is binomial and informs the respondents are legally married or not. Thus, the generalized linear model is not just a linear regression, instead, it is going to be logistic regression. This model tries to determine what respondents' characteristics indicate their marriage's history. Also, we want to see do the respondents' age, their feeling of their own lives and their self-rated health bring influence on the odds of having married legally or not. The formula we illustrated here tells us the relationship between the response variable which is whether the respondent ever legally married and all explanatory variables which are respondents' age, the feelings about their life, and their self-rated health. p is the probability of the male respondent have married legally.

$$\begin{aligned}\log \frac{p_i}{1-p_i} =& \beta_0 + \beta_1 x_{age,i} + \beta_2 x_{feelings life,i} - \beta_3 x_{self_rated_health_2,i} \\ & - \beta_4 x_{self_rated_health_3,i} + \beta_5 x_{self_rated_health_4,i} + \beta_6 x_{self_rated_health_5,i} + \epsilon_i\end{aligned}$$

The estimated coefficient β_1 and β_2 represent the change in log odds for every one unit increase in age and level of self-feeling of life. In other words, β_1 means when male respondent's age increases 1 year, the odds that they have married legally increase by 0.080%. The level of respondents feeling of their own life is ranked from 1 to 10, thus, β_2 means when male respondent's feeling of their life increases by 1 unit, their probability of having married legally increase 0.133%. We notice that the respondents' answers to self-rated health are categorical. Thus, we classify the self-rated health into 5 levels which are Excellent, very good, good, fair, and poor. We rank these five categories from 1 to 5 where 1 indicates poor and 5 indicates excellent. According to the rank, we can see that excellent is better than very good and very good is better than good, and so on. As a result, we use the dummy variable coding with the explanatory variable self-rated health to help us build the model. The first self-rated health category as poor is a reference category that does not require a variable. The estimated coefficient β_3 of the second self-rated health category is -0.240 illustrates that there is a decrease compared to rank 1 which is poor. Similarly, the estimated coefficient β_3 of the second self-rated health category is -0.247 tells us that there is a bigger decrease compare to rank 1. This concludes these two self-estimated feelings of life decrease the odds that the male resident in BC has legally married. Thus, the coefficient for the dummy variables for the ranks, β_3, β_4, β_5, and β_6 are not actual coefficient for that rank itself, it is the difference between that rank of the self-estimated feeling of life and the feeling of poor.

6 Result

In summary, the table above shows that participants who are in younger age might rate themselves higher health levels, which means age may have a positive relationship between health levels. For another table, we see that people with a worse feeling about life would rate themselves lower health levels, which means there is also a positive relation between them. The histogram of age shows participants who are 80 years old takes the biggest quantity of all participants. Moreover, the histogram of feelings about life tells us most participants are satisfied with their life. The scatter plot shows for people who rated their health levels as "poor", they intend to have a higher age than other participants. Also, people generally consider themselves healthy, except for ager people who self-rated themselves "Poor" health level. Furthermore, we established a logistic regression model and aimed to estimate the coefficient for the linear system of the relationship between the response variable which is whether the respondent ever legally married and all explanatory variables which are respondents' age, the feelings about their life and their self-rated health. The research result is whether males in BC have married or not is affected by their age and feelings of life and self-rated health.

7 Discussion

This model could predict the possibility of whether a male respondent has ever legally married or not in BC by given the corresponding respondent's age, self-rated health, and their feeling of life. The table of the summary of model estimates shows how each explanatory variable which is which are respondents' age, the feelings about their life, and their self-rated health brings influence the response variable which is whether

the respondents ever legally married. It is obvious that respondents' age and their feeling of life increase the possibility of having married. However, the effect of their self-rated health on whether the respondents ever legally married might be not very clear. Particularly, when male respondents think their health is fair and good, their odds of having married legally is lower than the respondents who think their health is poor, however, when male respondents think their health is excellent and very good, their odds of having married legally is higher than the respondents who think their health is poor. It may be caused by the influence of other factors. In general, age and feelings of life and self-rated health increase the possibility of having married for males in BC. Therefore, whether males in BC have married or not is affected by their age and feelings of life and self-rated health.

8 Weaknesses

Firstly, the most apparent limitation of our analysis is lurking variables. Even we use the logistic regression model to predict the married marital status among males in BC, we should still consider their financial status, ever second marriage, they have children or not, since those factors significantly affect the self-rated health. Moreover, some questions like "feelings about life" are considered inaccurate, since the participants are easy to alter their choices depending on their current moods, which could affect the true outcome of this study.

9 Next Steps

For future work, we can consider more lurking variables, like "financial status, ever second marriage, they have children or not", and make other regression models to predict the married marital status. Then, we can compare with our past work to find the impact of each factor. Secondly, making a residual plot of the models to see if there is any limitation of data. Finally, we would find effective methods to decrease the influence of outliers on the result.

References

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics.*

Canada, Statistics. 2017. "General Social Survey - Family (Gss)." https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey.

Clemens, Jason, and Sasha Parvani. 2019. "Canada Must Prepare for Our Aging Population: Op-Ed." *Fraser Institute*. https://www.fraserinstitute.org/article/canada-must-prepare-for-our-aging-population.

DePaulo, Bella. 2019. "The Single Life May Get Even Better with Age." *Psychology Today*. https://www.psychologytoday.com/us/blog/living-single/201901/the-single-life-may-get-even-better-age.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*.