What Improves People's Feeling of Life?

Boyuan Cui, Zhaocheng Li

Oct 19, 2020

Abstract

The report gives a brief analysis on what factors improve people's life satisfaction. Specifically, we explored the impact of marital status, education level, individual income, self rated health, and average working hours on the well-being of individuals based on the 2017 General Social Survey data of Canada. The report analyzes the correlation through Bayesian linear regression and uses Gibbs sampling to estimate the model parameters. We discovered that health, income, and certain marital status (married, living common law, and divorced) have a positive influence on the well-being, while education has a negative impact. Working hours and other marital status (widowed and separated) has no significant influence on the well-being.

Introduction

Starting from 1985, Statistics Canada has conducted General Social Survey (GSS) every year to collect cross-sectional data on non-institutionalized Canadian citizens over 15 years old to address and discover new social issues. One of the primary objective of GSS is to monitor factors that affect Canadians' well-beings of that specific year [1]. We believe that an important indicator of a person's well-being is his/her own attitude towards life. If we are aware of what aspects promote or damage one's feeling of life, the government and any social agencies will have better insights on what they should do to make the society better off. We believe it is important to know which factors are influential on people's feeling of life as a whole, and we may find some unexpected discoveries.

The following sections will specifically explore the impact of marital status, education level, individual income, self rated health, and average working hours on one's feeling of life, and we will apply a Bayesian linear model to analyze the relationship.

Data

Data Extraction

We first obtain the GSS data file from the CHASS website ('AAIpsmGG.csv'), and clean the data with code from 'gss_cleaning.R' (written by professor Rohan Alexander and Sam Caetano) to make the raw data more interpretable. The data file after cleaning is 'gss.csv'

Then, we select our interest variables–marital status, education level, individual income, family income, self rated health, average working hours, and one's feeling of life–from the cleaned data, and we remove every case with NA or "Don't know" responses from the dataset. For the convenience of modeling, we will convert some categorical data to numeric variables. We will briefly explain each variable in the following section.

Below is a glimpse of the final dataset:

Table 1: Glimpse of Dataset

feelings_life	marriage	education	health	income_respondent	work_hours
8	Single, never married	2	5	2	3
10	Married	3	3	1	5
8	Living common-law	4	3	1	3
8	Single, never married	2	4	1	3
10	Single, never married	1	5	1	2
6	Single, never married	5	4	2	5

Variable Explanations

Feeling of life (code in raw data: SLM_01, name in 'gss.csv': feelings_life) is a continuous numeric data ranging from 0 to 10 that indicates the respondent's feeling of life as a whole. 0 means very dissatisfied and 10 means very satisfied (reference from data).

Marital status (code in raw data: MARSTAT, name in 'gss.csv': marital_status) is a categorical data indicating whether the respondent is divorced, living common-law, married, separated, single, or widowed [1]. We will not convert this data to numeric, since it is not intuitive to differentiate marital status into different levels.

Education level (code in raw data: EHG3_01B, name in 'gss.csv': education) is a categorical data that indicates the highest education the respondent acquires [1]. Because the question asks for the highest education level achieved, it is reasonable to use numeric data to represent different levels of education. The variable after transformation will range from 1-7, representing from the lowest to the highest education level (we use the order of education level from the original GSS). Below is a reference table:

Table 2: Education Level Reference

Education Level	Numeric Rank
Less than high school diploma or its equivalent	1
High school diploma or a high school equivalency certificate	2
Trade certificate or diploma	3
College, CEGEP or other non-university certificate or diploma (other than trades certificates or diplomas)	4
University certificate or diploma below the bachelor's level	5
Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	6
University certificate, diploma or degree above the bachelor's level	7

Individual income (code in raw data: TTLINCG2, name in 'gss.csv': income_respondent) is a categorical data that indicates the income level of respondent before tax and excluding any capital gains or losses [1]. Income is inherently a numeric data, so we will transform these two variables to numeric data ranging from 1 to 7 which represent the income level. Below is a reference table:

Table 3: Income Level Reference

Income Level	Numeric Level
Less than $25,000	1
$25,000 to $49,999	2
$50,000 to $74,999	3
$75,000 to $99,999	4

Income Level	Numeric Level
$100,000 to $ 124,999	5
$125,000 and more	6

Self rated health (code in raw data: SRH_110, name in 'gss.csv': self_rated_health) is a categorical data that indicates what the respondent think his/her health level is [1]. We will also transform this variable to numeric data with range from 1-5. Below is the reference table:

Table 4: Health Level Reference

Self Rated Health Level	Numeric Level
Poor	1
Fair	2
Good	3
Very good	4
Execellent	5

Average working hours (code in raw data: UHW_16GR, name in 'gss.csv': average_hours_worked) is a categorical data that indicates how many hours the respondent works a week on average [1]. We will convert the variable to numeric with range from 1-5. Below is the reference table:

Table 5: Work Hour Level Reference

Weekly Work Hour Level	Numeric Level
0 hour	1
0.1 to 29.9 hours	2
30.0 to 40.0 hours	3
40.1 to 50.0 hours	4
50.1 hours and more	5

Why do we choose these variables?

The feelings about life or life satisfaction has been a popular topic in the society and should be a major concern of the governments, so we think it is a profound topic to be analyzed and discussed. We try to figure out some key factors that might affect people's feelings about life and discover how exactly they would change the feelings. Therefore, we choose feelings about life to be the response variable, then we browsed the data set and we found some potential explanatory variables based on our knowledge. For example, people usually will be more satisfied with their life when they are in a better health condition; also the people who are single might have different feelings about life from the people who are married. Because life consists of many aspects so we chose five different key factors from five different aspects, and the results would be relatively representative and significant.

Potential Drawback

For the health variable, we use the self rated health from GSS, but this does not necessarily represent the actual health condition of respondent. The respondent himself/herself might not be a physician and could have wrong understanding of his/her actual health condition.

Model

We will use a Bayesian linear model.

In other words, we are trying to find a linear formula that describes the relationship between our explanatory variables and the response variable:

$Y_{feeling} = \beta_0 + \beta_1 \cdot I_{divorced} + \beta_2 \cdot I_{CommonLaw} + \beta_3 \cdot I_{married} + \beta_4 \cdot I_{separated} + \beta_5 \cdot I_{widowed} + \beta_6 \cdot X_{education} + \beta_7 \cdot X_{health} + \beta_8 \cdot X_{income} + \beta_9 \cdot X_{hours}$

where,

$Y_{feeling}$ is the model's predicted feeling of life as a whole

I_i represents an indicator variable of marital status that can only take value of 0 or 1

For $I_{divorced}$, 1 means the person is divorced and 0 means not.

For $I_{CommonLaw}$, 1 means the person is living in common law and 0 means not.

For $I_{married}$, 1 means the person is married and 0 means not.

For $I_{separated}$, 1 means the person is separated and 0 means not.

For $I_{widowed}$, 1 means the person is widowed and 0 means not.

$X_{education}$ is a variable with range from 1 to 7, indicating the education level.

X_{health} is a variable with range from 1 to 5, indicating the self rated health level.

X_{income} is a variable with range from 1 to 6, indicating the respondent income level.

X_{hours} is a variable with range from 1 to 5, indicating the average weekily working hour level.

For the model to make sense, if one marital status indicator variable I_i is 1, all other I_i should be 0. The variable I_{single} for Single is missing because we use it as a reference variable. In other words, the coefficient β_1 of $I_{divorced}$ represents how much the respondent's feeling of life as a whole will change, comparing to being Single, if the respondent is divorced, and the same applies to other I_i.

We will find βs that give the best fitted model.

We will assume that all these variables are independent. The difference between a simple linear regression model and a Bayesian one is that, in Bayesian approach, we see the parameters (i.e. βs) as random variables instead of fixed constant, and the probability distribution of βs is called a prior distribution.

$$prior = P(\beta_1, ..., \beta_9)$$

The objective of Bayesian is to find the conditional distribution of β given the sample data (i.e. data in Table 1), called posterior distribution.

$$posterior = P(\beta_1, ..., \beta_9 | data)$$

By Bayes' rule, the posterior distribution is proportional to the product of likelihood, which is the joint density distribution of all explanatory variables, and prior.

$$posterior \propto prior \cdot likelihood = P(\beta_1, ..., \beta_9) \cdot P(data | \beta_1, ..., \beta_9)$$

To calculate likelihood, we will just multiply the density distribution of all variables since we assume all variables are independent.

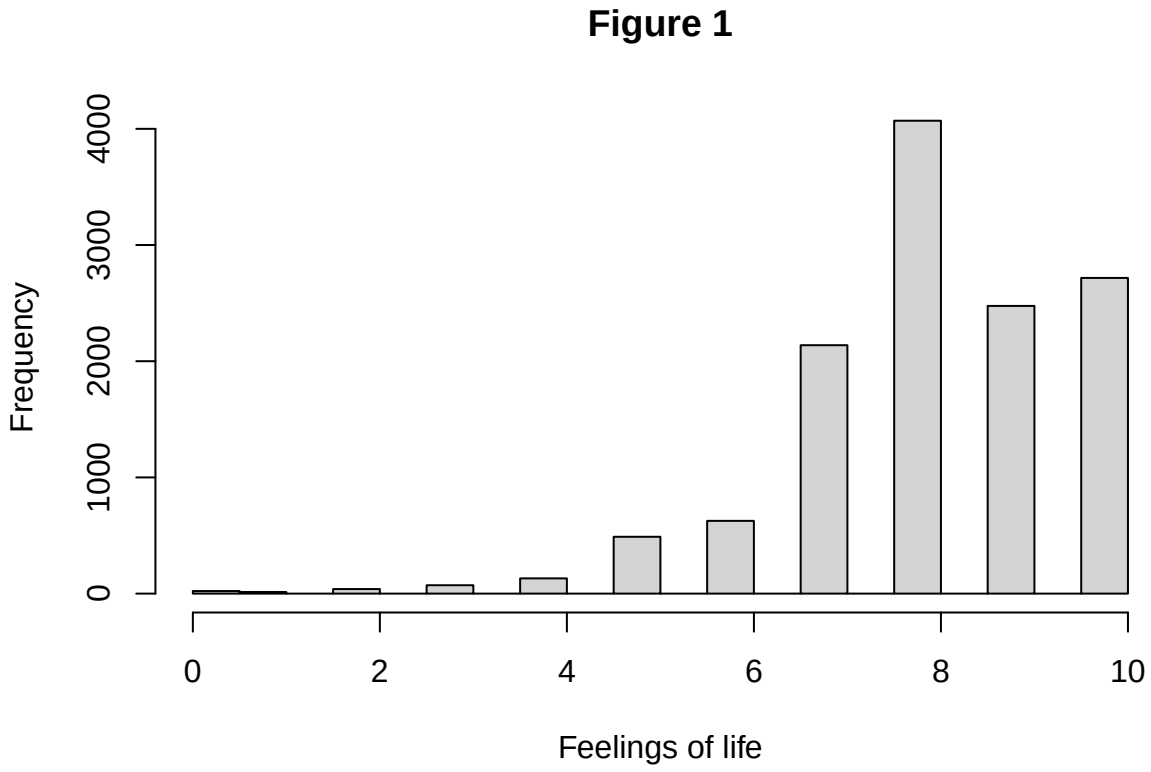
We decide to choose Bayesian because it gives a more compact confidence/credible interval than the simple linear model does [2]. With the same level of confidence, if the interval is smaller, we could give a more precise estimate of the parameter.

We will use the technique of Gibbs Sampling to estimate our coefficients. Since we have multiple coefficients, the distribution of one β will condition on the other βs, and Gibbs sampling provides an iterative algorithm that helps us to estimate each coefficient [3].

Results

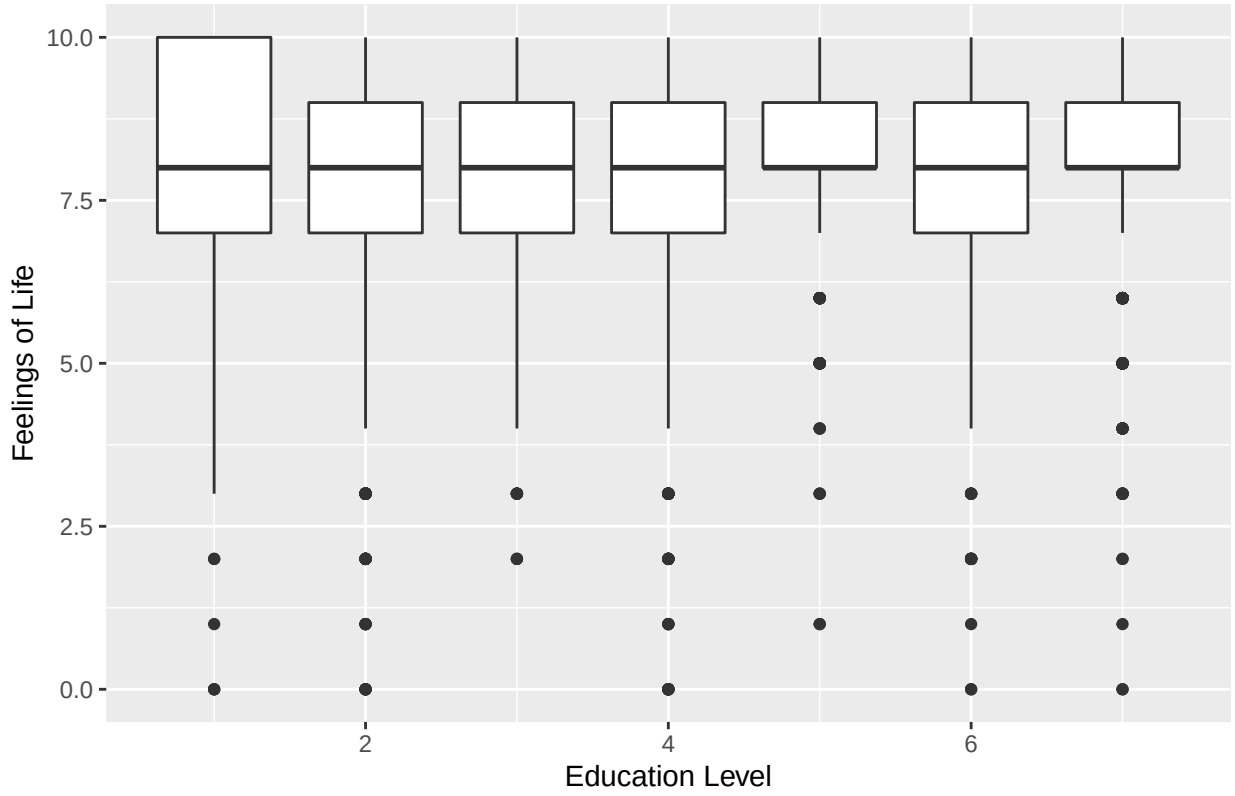
Data Overview

Below is a histogram that gives a brief overview of Canadians' feelings of life as a whole in 2017:



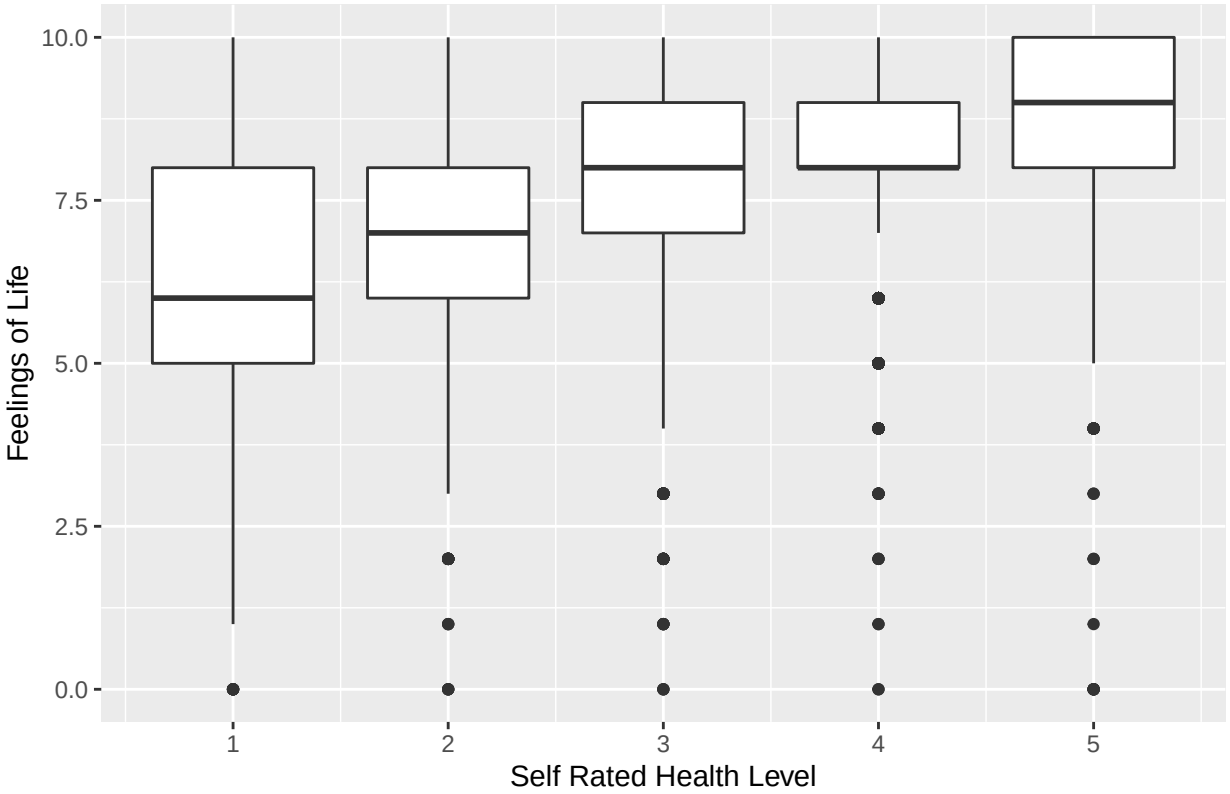
The distribution in Figure 1 is left skewed with a mean of 8.13, so we can say that the overall well-being of Canada in 2017 is quite good.

Figure 2



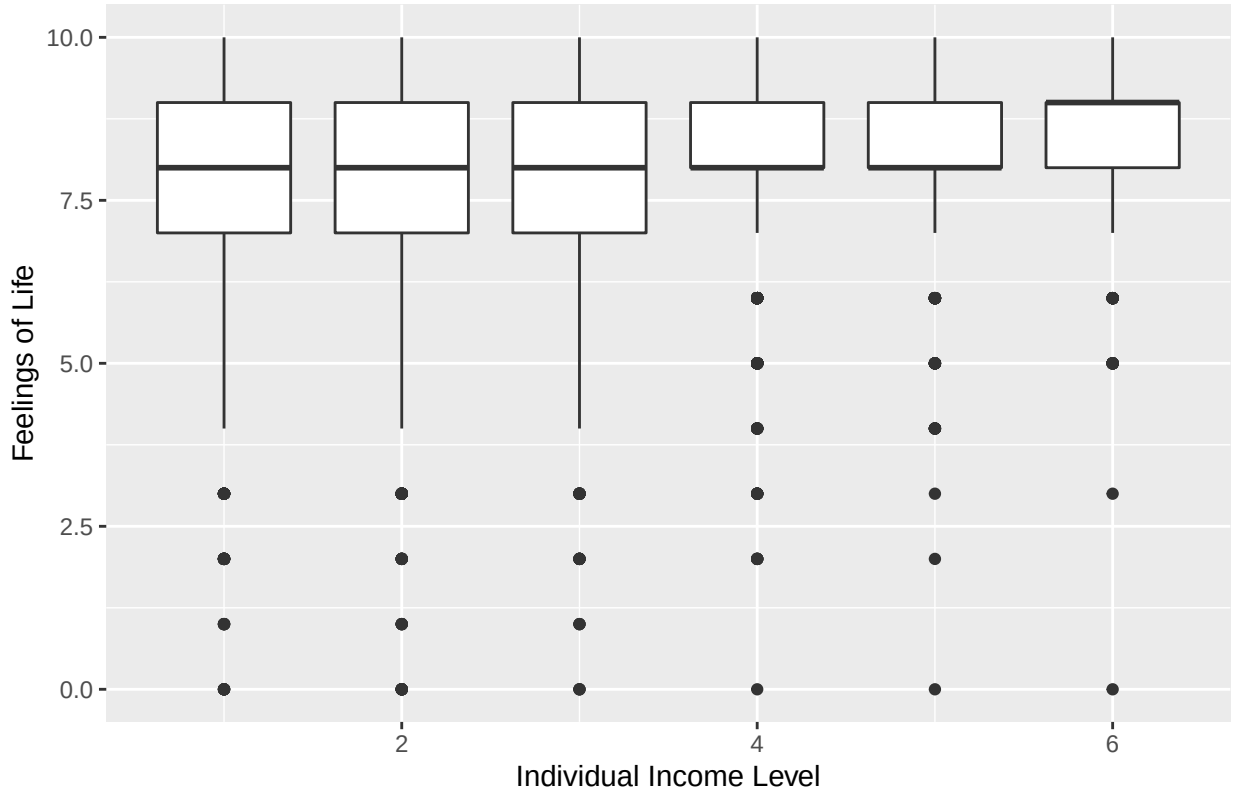
Above is a boxplot of education level versus feeling of life. It is not obvious to see whether there is a relationship; we will need the linear model to check the correlations.

Figure 3



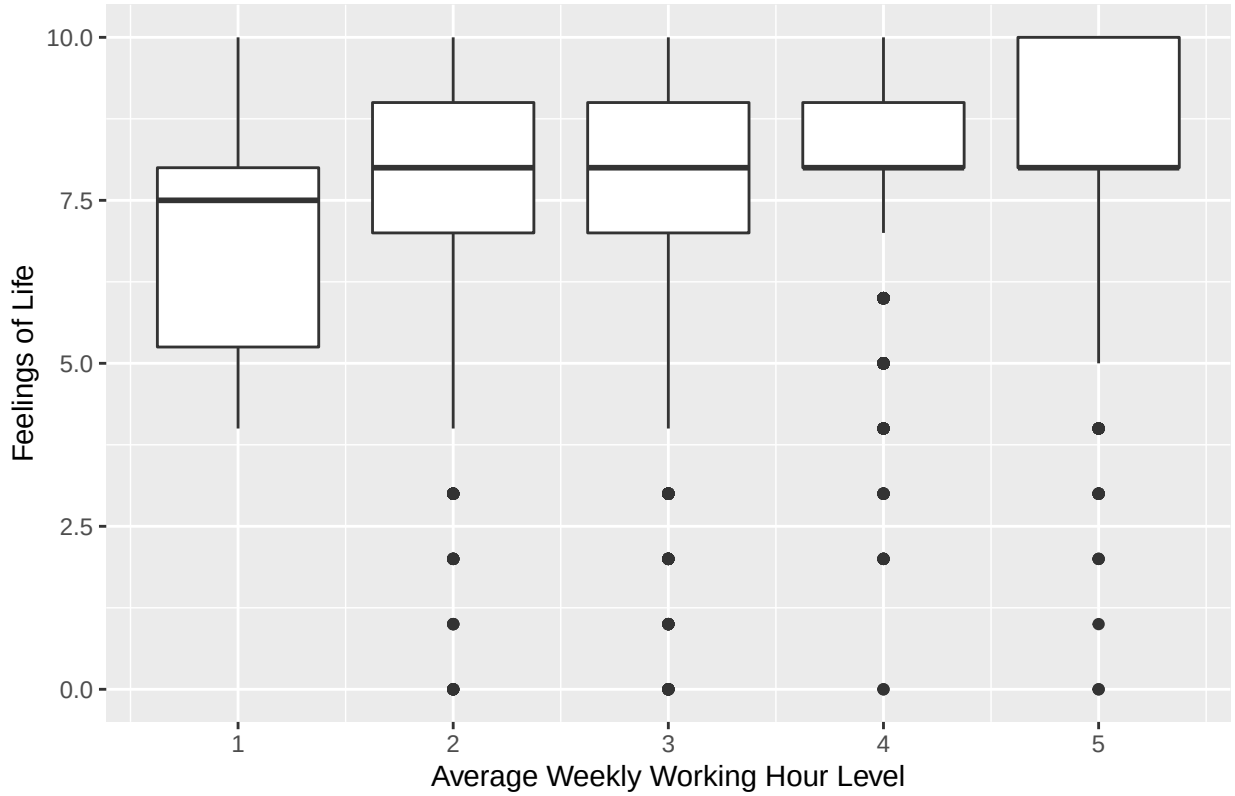
Above is a boxplot of self rated level versus feeling of life. We can see that there is positive correlation between self rated health and feeling of life, but we still need the model to verify this.

Figure 4



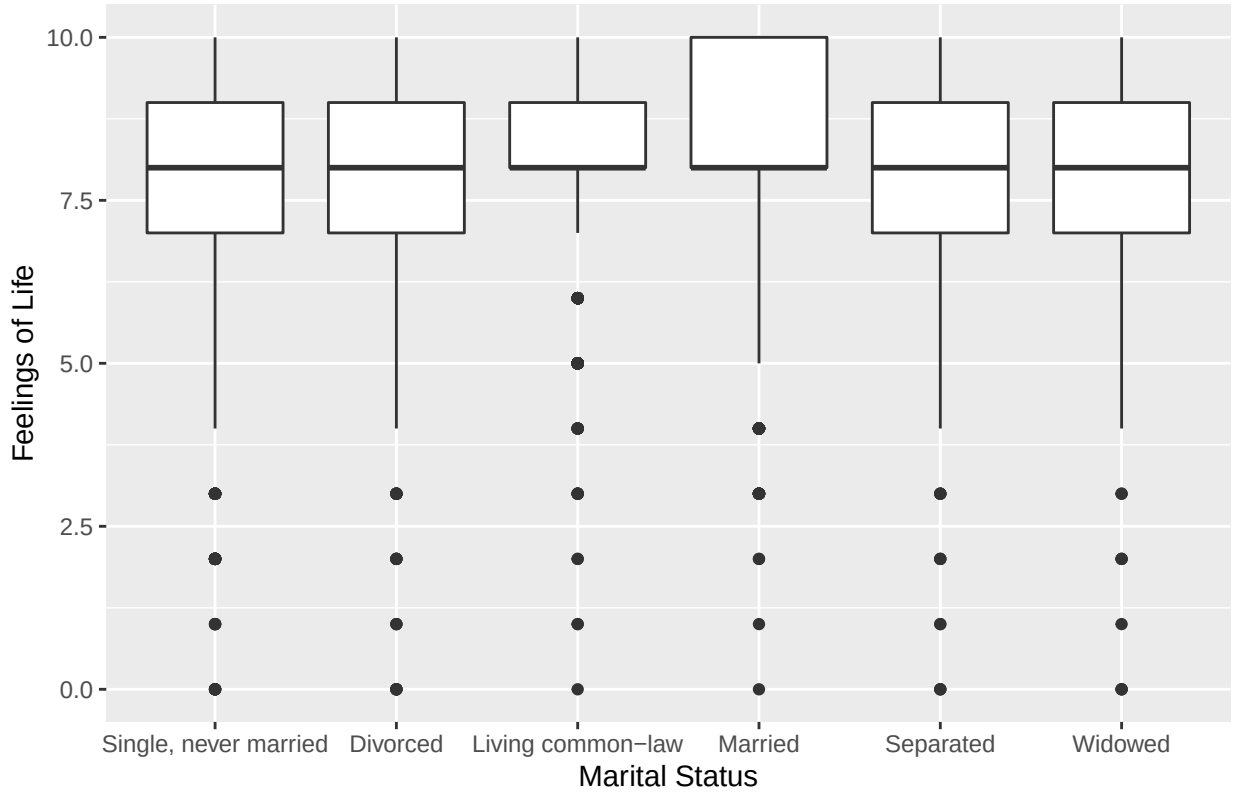
Above is a boxplot of respondent income versus feeling of life. Just by looking at the graph, we cannot see a relationship between these 2 variables.

Figure 5



Above is a boxplot of weekly average working hours versus feeling of life. Just through the plot, we might propose that longer work hours is correlated with more satisfied feeling of life; we will need the model to verify this.

Figure 6



Above is a boxplot of marital status versus feeling of life. Just by looking at the graph, we cannot clearly see a relationship between these 2 variables.

Regression Result

Below is a summary of our result Bayesian linear model:

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: feelings_life ~ marriage + education + health + income_respondent + work_hours
##    Data: gss_numeric (Number of observations: 12794)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                    5.73      0.08     5.58     5.88 1.00     6905
## marriageDivorced             0.16      0.05     0.06     0.26 1.00     4634
## marriageLivingcommonMlaw     0.52      0.04     0.43     0.60 1.00     4635
## marriageMarried              0.73      0.03     0.67     0.79 1.00     4095
## marriageSeparated           -0.10      0.07    -0.25     0.04 1.00     5385
## marriageWidowed              0.01      0.09    -0.16     0.18 1.00     6403
## education                   -0.06      0.01    -0.07    -0.04 1.00     7646
## health                       0.54      0.01     0.52     0.57 1.00     8710
## income_respondent            0.04      0.01     0.02     0.06 1.00     6006
```

```
## work_hours                   0.02      0.02    -0.01     0.05 1.00     6828
##                           Tail_ESS
## Intercept                     3241
## marriageDivorced              3340
## marriageLivingcommonMlaw      3196
## marriageMarried               3473
## marriageSeparated             2927
## marriageWidowed               3262
## education                     3334
## health                        3065
## income_respondent             2952
## work_hours                    3298
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     1.39      0.01     1.37     1.40 1.00     9193     2954
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

We will focus on the table under 'Population-Level Effects'.

From the results of Bayesian linear regression, firstly we will check whether there is any correlation between response variable and each independent variable. We will use hypothesis testing. The null hypothesis here is that there is no relationship between feelings about life and each factor. If the hypothesis value of the null, which is 0, is in the 95% confidence interval, then the we will have no evidence to reject the null hypothesis; if the confidence interval does not contain the null hypothesis, we could reject the null hypothesis.

The lower and higher bound of 95% confidence interval can be found under the 'l-95% CI' column and 'u-95% CI' column, respectively.

As we can see, the 95% confidence intervals for marital status of separated ($[-0.25, 0.04]$), widowed ($[-0.16, 0.18]$) and work hours ($[-0.01, 0.05]$) do contain 0. Hence we could conclude that the feelings about life does not correlate with each of these three variables.

Secondly, if the null hypothesis is rejected, we will analyze the effects of predictor variables on the outcome variable by checking the estimated values of β.

The estimated value of β can be found under column 'Estimates', where each value, from top to bottom, represents estimated value of $\beta_0, ..\beta_9$, respectively.

From the model, it is clear that comparing to participants who are single, those participants who got divorced, living common-law and married have better feelings about life, with 0.16 units, 0.52 units, 0.73 units increment respectively. Moreover, respondents are 0.54 units more satisfied with life as their health condition moves forward to the nearest better level(such as from "Good" to "Very good", reference from Table 4), and the rise from lower income range to the nearest higher income range (e.g. "Less than $25,000" to "$25,000 to $49,999", reference from Table 3) grows their life satisfaction by 0.04 units.

On the contrary, the level of education has a negative impact on the feelings about life. From the model, we discovered that if the educational level increases from lower level to the nearest higher level(e.g. the changes from "Less than high school diploma or its equivalent" to "High school diploma or a high school equivalency certificate", reference from Table 2), the feelings about life is predicted to decrease by 0.06 units.

The intercept will have no practical meaning here, because all variable ranges do not contain 0.

Discussion

As we can see from our model, a good health condition greatly promotes personal well-beings in Canada, and it has greater positive impact on well-beings than the individual income has. This result implies that in order to increase the happiness of the society, the government should focus more on public health policies and personal medical support, etc.

Additionally, from the model, we find out that getting married and living common-law also contributes largely to Canadian's feeling of life. This is intuitive since we expect people to be happier when they are in a relationship. This result may imply that the government should improve the common law or simplify the process of marriage registration.

We also discover from the model that higher education level has a tiny negative impact on people's feeling of life. A possible explanation is that as people obtain more knowledge, they might worried about more things. For example, a construction worker may feel very satisfied with his/her job while an economic professor is constantly worrying about an upcoming recession. This does not necessarily mean that government should discourage people from increasing their education level. In fact, a higher education level might relate to a higher income, and by our model, income has a positive influence on the well-being.

One interesting discovery is that, comparing to single respondents, divorced people have a more positive feeling of life. Intuitively, we often relate divorce to grief and distress, but the data from GSS tells a different story. Based on this result, the government may need to change its view on divorce, and consider to modify terms in the marriage laws to address this social issue.

Weaknesses

When we fit our model, we assume that all variables are independent, but that is not necessarily true. The education level and income may have relations because a good education background possibly helps people to get a high-waged job. Average work hours could also relate to health since if a person works a lot, his/her health condition could be affected.

Next Steps

We could further investigate why marital status of divorce has a positive impact on people's feeling of life, and it may require another linear model to discover potential reasons.

References

[1] Statistics Canada. 2017. General Social Survey on Family (Cycle 31).

[2] Educational Reseatch Techniques. 2018. Linear Regression vs Bayesian Regression. Retrieved October 18, 2020 from https://educationalresearchtechniques.com/2017/10/18/linear-regression-vs-bayesian-regression/

[3] Ilker Yildirim. 2012. Bayesian Inference: Gibbs Sampling. Retrieved October 18, 2020 from http://www.mit.edu/~ilkery/papers/GibbsSampling.pdf

Appendix

Github repository of this report: https://github.com/cuiboyuan/GSS2017_Life_Satisfaction_Analysis