The relationship between feelings life and family and health

Linxia Li (1005715488), Leyi Wang (1006318682), Xingnuo Zhang (1006145306), Yanlin Li (1003770305)

October 23, 2020

Abstract

In our research, we are interested in analyzing how family (marital status and total children) and health (self-rated health and self-rated mental health) can affect respondents' feeling of life in the GSS dataset as feelings_life is a good indicator to reflect citizens' standard of living. By making assumptions and fitting a multiple linear regression model, we found that there exists a positive correlation between explanatory variable and response variable. To illustrate, in the family aspect, along with the increase in stability of marital status and number of children, the index of feeling of life increases. Similarly, in the health aspect, the physical and mental health enhances the quality of life.

Introduction

In this report, we focus on General Social Survey (GSS). The goal of GSS is to monitor the overall standard of living and well-being of Canadian by analyzing the social trend and citizens' thoughts. Besides, it acts as a guide and provides information of current interest on social issues. We chose this data because by analyzing this data, we can conclude how some features of interest can affect Canadian's well-being and what can make our lives better.

We identify respondents' feelings of life as our major interest, because we think it is a representative feature of Canadian's living standard and well-being. Thus, choosing this variable is consistent with our goal mentioned above.

In this report, we will explore how a respondent's feelings of life can be affected by two of the major aspects of his (or her) life: family and health. We identified two features for each aspect. For family, we consider marital status and total children as the representative parts. In case of health, self rated health and self rated mental health are the chosen features.

We use linear regression to analyze our data and made the following inferences:

1. In family aspect, if people have more stable conjugal relationship and children, they tend to have a better feeling of life.
2. In health aspect, if people have a physical and metal health condition, they also tend to have a better feeling of life.

Here is a brief description about each part of our report:

1. **Data:** We will discuss the survey and the data, the key features strengths and weaknesses about them and some plots and raw data.
2. **Model:** We will establish our variables and present the relationship between the variables we are going to use. Besides, we will choose the model and explain it, discuss the features and the final model and do diagnostic checks.
3. **Results:** We will present our results and explain them.

4. **Discussion:** We will talk about the survey, data and the model, identify strengths and weaknesses.

The cost of carrying out this survey is absolutely high, no matter in time and energy or in financial aspect.

Data

Introduction of the Survey Data

The General Social Survey (GSS) program, established in 1985, conducts telephone surveys across the ten provinces. The GSS is recognized for its regular collection of cross-sectional data that allows for trend analysis, and its capacity to test and develop new concepts that address current or emerging issues. In the GSS dataset, it contains 20602 observations and 81 variables.

The participants took this survey by Computer Assisted Telephone Interview, with their selected language.

The target population (The set of all the units covered by the main objective of the study): All persons whose age is equal or above fifteen in Canada, excluding residents living in Yukon, Northwest Territories and Nunavut and full-time residents of institutions.

Sampling frame (A source material or device/list from which a sample is drawn): The sampling frame includes two components. One is the lists of telephone numbers in current use which can be found at Statistics Canada; The other is the Address Register (AR), which is a list of dwellings within the ten provinces. AR is able to link all available telephone numbers with the same valid address and 86 percent is linked successfully by AR.

The frame population: (The set of all units covered by the sampling frame) The frame population is all the targeted population that can be accessible by the sampling frame listed above.

Sample: (The population represented by the survey sample) The randomly selected persons to participate in the telephone interview.

Sampling method: In order to obtain the survey sample, they applied a method called Stratified Random Sampling, which divides the population into smaller sub-groups by homogeneous characteristic known as strata. This method ensures the equally likely probability of being selected from the entire population. By this method, they divided ten provinces into strata by geographic areas. There exist 27 strata in total. In each stratum, they used Simple Random Sampling without replacement. Specifically, this method assigned a sequence of number in the stratum, and used random number generator to do the sample selection. After that, they obtained the 2017 GSS dataset.

Non-response Problems and solutions: Targeting on those people who at first refused to response, they would re-contact them again up to two more times and try to motivate them to participate by explaining the importance of this survey. Besides, if the timing of the interviewer's call was inconvenient, they would make an appointment with them and call back. If there was no one home, they would call backs until they response.

Strengths and Weaknesses

Strengths

The strength of the data: Based on the large sample size, we are able to analyze the relationship among different variables and have an accurate inference to support assumptions.

The strength of the survey: It involves many aspects of living conditions such as marital status, health condition, and income. Therefore, it provides an overview about the basic living conditions of people aged 15 and over in Canada.

Weaknesses

The weaknesses of the data: There exists lots of non-response data (NA). After we filter out those data, it does not play an essential role in influencing the response variable. To be specific, it reduces sample size, which decreases the precision of estimators and increases the standard error.

The weaknesses of the survey:

1. The question “Using a scale of 0 to 10 means”Very dissatisfied" and 10 means “Very satisfied”, how do you feel about your life as a whole right now?". For the option of this question, 0 means very dissatisfied, 5 means generally satisfied and 10 means very satisfied. However, some people who were not satisfied with their life might select 0, 1 or 2 so that this choice makes the resulting data become less accurate.
2. Some questions include privacy issues so that a number of respondents are not willing to disclose their personal information or provide inaccurate data instead. Either one of these situations may influence the accuracy of our analysis and conclusion.
3. They did not introduce weights to each groups of people, such as gender. Because different groups may have differences between each other, there may be biases that can make the sample not representative to the whole population.

Feature selection

Here are the features that we selected into our dataset.

Sample composition: age, sex

Response variable: feelings of life.

Features: (See Appendix for more about how we select features)

1. Family: marital_status, total_children
2. Health: self_rated_health, self_rated_mental_health

We reject features based on the following criterion:

1. Too many NAs: If there are too many NAs in this feature, we reject it for the convenience of our analysis. Rejecting these features can also avoid biases caused by non-response.
2. Replication: For our linear regression model, we should make sure that each features are independent. Thus, we choose the most representative ones among a certain type.
3. Not correlated with response variable: For the simplicity of our model, we tend to choose the features that are most correlated with the model.

(See Appendix for more about how we select features)

Cleaning the data

We clean the data using the package `tidyverse` and removed all the NAs in the cleaned data. Here is the first few rows of the data we will use. (Table 1)

Table 1: Table 1 - First few rows of data

feelings_life	age	sex	marital_status	total_children	self_rated_health	self_rated_mental_health
8	52.7	Female	Single, never married	1	Excellent	Excellent
10	51.1	Male	Married	5	Good	Good
8	63.6	Female	Married	5	Very good	Good
10	80.0	Female	Married	1	Very good	Very good

feelings_life	age	sex	marital_status	total_children	self_rated_health	self_rated_mental_health
8	28.0	Male	Living common-law	0	Good	Good
9	63.0	Female	Married	2	Excellent	Very good

Here are two summary about the numerical (Table 2) and categorical (Table 3) data that we will use.

Table 2: Table 2 - Summary of numerical data

feelings_life	total_children
Min. : 0.000	Min. :0.000
1st Qu.: 7.000	1st Qu.:0.000
Median : 8.000	Median :2.000
Mean : 8.095	Mean :1.674
3rd Qu.: 9.000	3rd Qu.:3.000
Max. :10.000	Max. :7.000

```
## Warning in kable_pipe(x = structure(c("Maritial status", "", "Self rated
## health", : The table should have a header (column names)
```

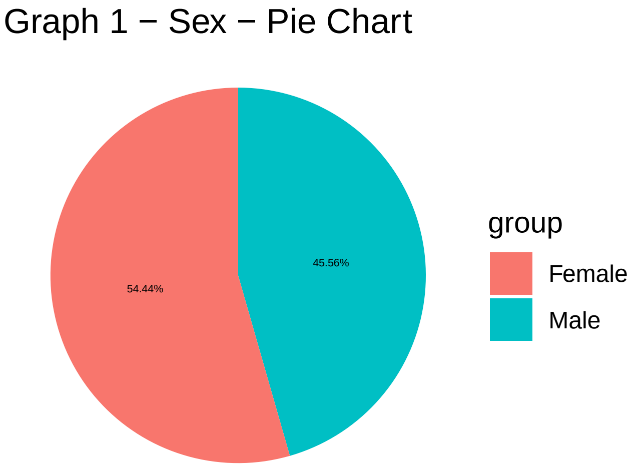
Table 3: Table 3 - Counting categorical data

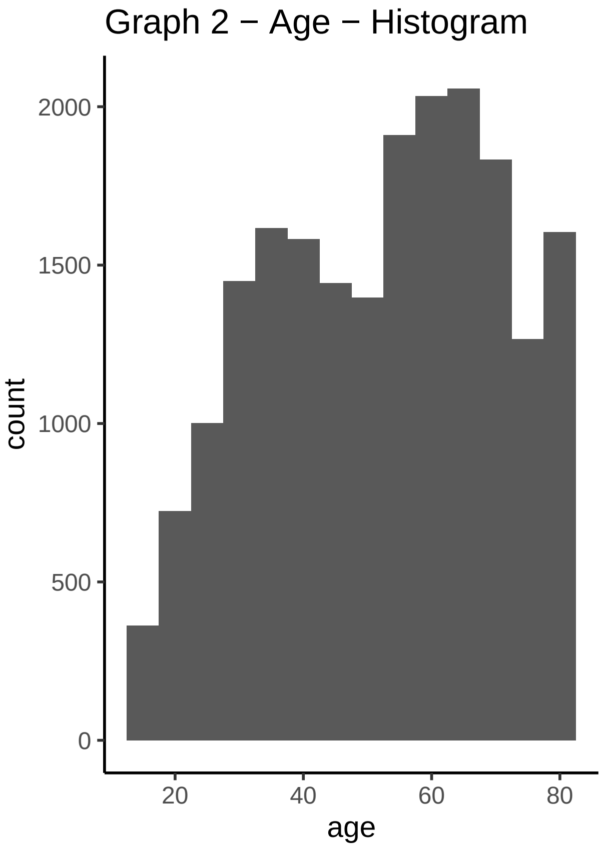
Maritial status	Divorced	Living common-law	Married	Separated	Single, never married	Widowed
	1737	2056	9360	627	4653	1850
Self rated health	Don't know	Excellent	Fair	Good	Poor	Very good
	47	4352	2037	6086	788	6973
Self rated mental health	Don't know	Excellent	Fair	Good	Poor	Very good
	40	6039	1272	5736	312	6884

Presenting the data

Basic composition of the sample

Here are two plots that can depict the basic composition of the sample: The sex (Graph 1) and the age (Graph 2).





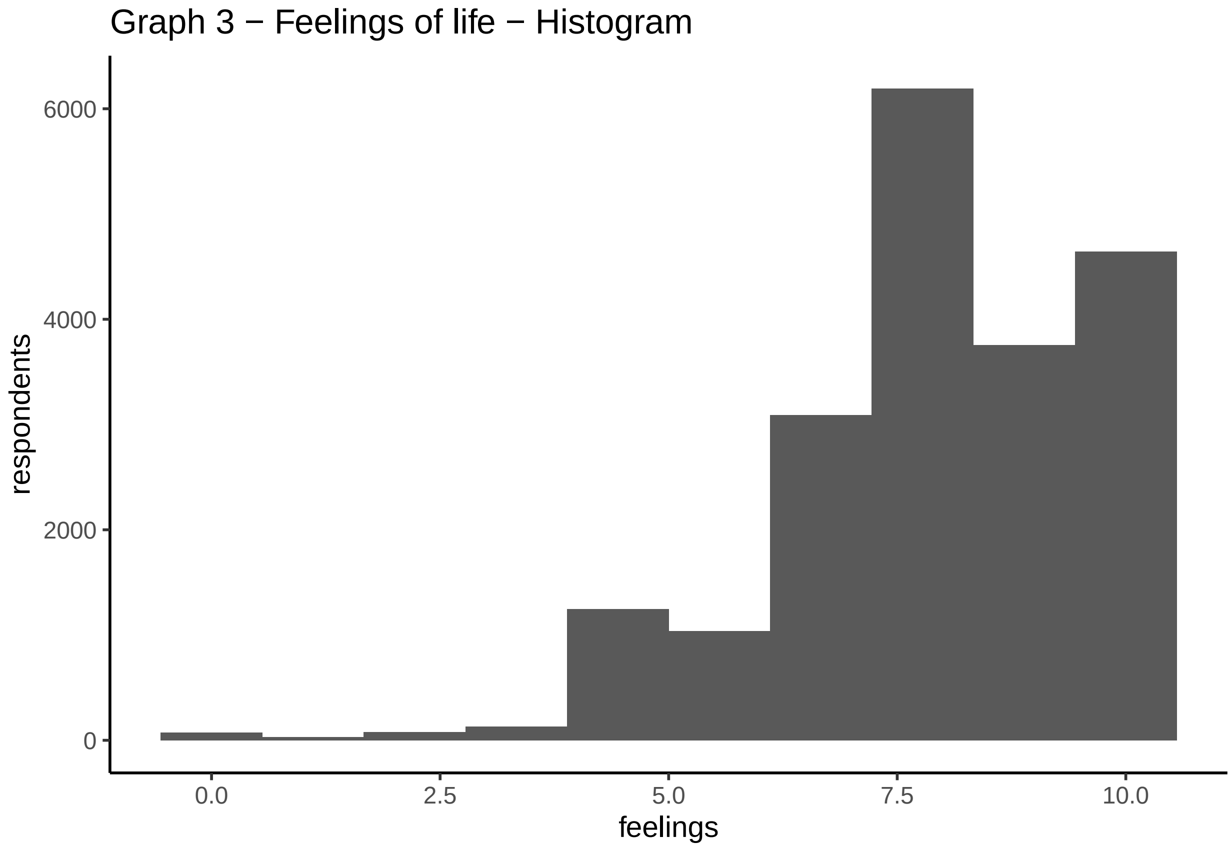
Inference:

1. Sex: There are slightly more female than male, but the difference is not large.
2. Age: The histogram is left-skewed. So there are more middle-aged and elderly respondents than young respondents. However, it covers all the targeted age groups.

Thus this sample is valid for inferences.

Variable of Interest

Here is the Histogram for our variable of interest – Feelings of life (Graph 3).



Graph 3 − Feelings of life − Histogram

Inference: From the above histogram of feelings_life:

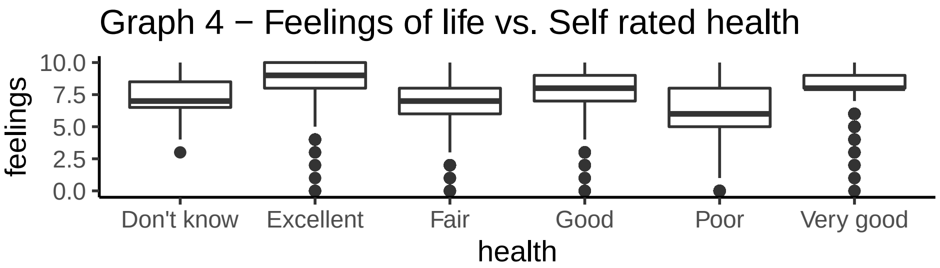
1. The distribution is left-skewed, single-peaked. The center of it might be 7.5 and the feelings_life of those non-institutionalized persons of 15 years of age and older surveyed, living in the 10 provinces of Canada is mostly concentrated between 6 and 10.
2. There are some outliers in this histogram which are representative for those small number of persons who had a lower satisfactory of life or those who misreported the data.

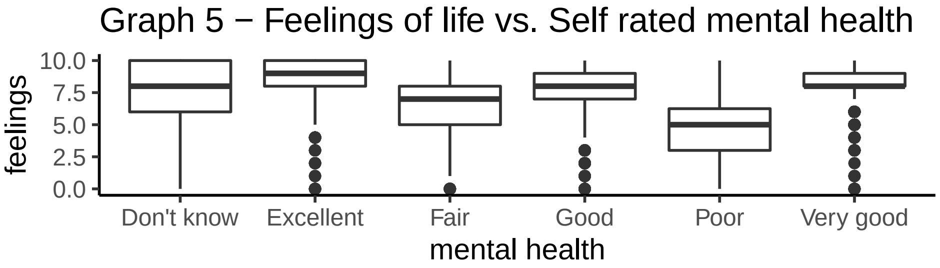
Model

Model selection

Relationships between variables

Health & feelings of life Below is the boxplots between health and feelings of life (Graph 4 and Graph 5)


Graph 4 − Feelings of life vs. Self rated health


Graph 5 − Feelings of life vs. Self rated mental health

Inference:

1. Self rated health:

a. The medians of boxplots (Excellent, Fair, Good) are in the middle so that the distributions of those three boxplots are symmetric. The medians of boxplots (Don't know, Poor, Very Good) are below the middle so that the distributions of those two boxplots are right-skewed.

b. The IQR (Interquartile range) of the boxplot (Poor) is the largest. The IQR of those four boxplots (Don't know, Excellent, Fair and Good) are approximately the same. The IQR of the boxplot (Very Good) is the smallest. However, some of them have some ourliers which are representative for those small number of persons who had a lower satisfactory of life.

2. Self rated mental health:

a. The medians of boxplots (Don't know, Excellent, Good) are in the middle so that the distributions of those three boxplots are symmetric. The medians of boxplots (Fair, Poor) are above the middle so that the distributions of those two boxplots are left-skewed. The medians of the boxplot (Very Good) is below the middle so that the distribution of the boxplot is right-skewed.

b. The IQR (Interquartile range) of the boxplot (Don't know) is the largest. The IQR of those two boxplots (Fair and Poor) are approximately the same. The IQR of those two boxplots (Excellent and Good) are approximately the same. The IQR of the boxplot (Very good) is the smallest. However, some of them have some ourliers which are representative for those small number of persons who had a lower satisfactory of life.

Family & Feelings of life Below is the summary table for relationship between feelings of life and marital status (Table 4) and total children (Table 5)

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

Table 4: Table 4: Feelings of life & marital status

marital_status	mean_feelings_life	sd_feelings_life	min_feelings_life	max_feelings_life	median_feelings_life
Married	8.432799	1.413826	0	10	8
Living common-law	8.224222	1.427167	0	10	8
Widowed	7.954054	1.840752	0	10	8
Divorced	7.679332	1.909289	0	10	8
Single, never married	7.673114	1.749819	0	10	8
Separated	7.317384	1.993536	0	10	8

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

Table 5: Table 5: Feelings of life & total children

total_children	mean_feelings_life	sd_feelings_life	min_feelings_life	max_feelings_life	median_feelings_life
4	8.336842	1.616335	0	10	8
6	8.303371	1.635790	2	10	8
7	8.287293	1.627986	2	10	8
2	8.213608	1.599885	0	10	8
3	8.197483	1.643221	0	10	8
5	8.196078	1.756058	0	10	8
1	8.045439	1.629073	0	10	8
0	7.879955	1.670419	0	10	8

Inference:

1. Marital status:

The summary table above shows the mean, median, minimum, maximum and standard deviation of feelings_life from different kinds of marital status. We have known that mean or median is a measurement of the center of a variable so that I use arrange in R to arrange the mean_feelings_life of different kinds of marital status. We can find out that those people who have married had the highest mean_feelings_life and those whose marital status is separated had the lowest mean_feelings_life.

2. Total children

The summary table above shows the mean, median, minimum, maximum and standard deviation of feelings_life from different numbers of total children. We have known that mean of median is a measurement of the center of a variable so that I use arrange in R to arrange the mean_feelings_life of different numbers of total children. We can find out that those people who have 4 children had the highest mean_feelings_life and those who did not have children had the lowest mean_feelings_life.

Model identification

We use four variables such as marital_status, total_children, self_rated_health and self_rated_mental_health to fit a multiple linear regression model with feelings_life as a response variable.

Discussions of the four independent variables (features)

Marital_status: This is a categorical variable that has five types. We use marital_status rather than marital_status-groups since different types of marital status can have different influences on the feelings_life. From the above summary table we make, we can find out that different types of marital_status has various mean of feelings_life so that we should use it instead of the groups of it.

total_children: This is a numerical variable that is a data variable taking on any value within 0 and 7. We can not treat it as a categorical variable in our model since different numbers of children can have various influences on the feelings_life. If we treat it as a categorical variable, our predictions and the analysis of the model might be inaccurate.

self_rated_health and self_rated_mental_health: These are two categorical variables which measure the quality of the physical and mental health of a person.

Overall, we chose these four independent variables and use them in the model to predict feelings_life since all of them can influence the feelings life of a person. We chose feelings_life as a response variable since it can reflect the overall standard of living of a person. Besides, in the survey, the question related to it divided feelings_life into different levels so that we can have a better understanding of the living conditions of people in Canada by using it as the response variable and analysing it.

Reasons to choose multiple linear regression model

1. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line. Here, marital_status, total_children, self_rated_health and self_rated_mental_health are four independent variables and all of them can influence feelings_life which is a quantitative dependent variable. Therefore, we can use multiple linear regression model to estimate the relationship between them.

2. For the Bayesian model, it assumes the parameters we would like to explore follow some distributions. However, we did not explore the distributions of those four variables so that we can not use the Bayesian model.

3. For the logistics regression model, it is a statistical model that in its basic form uses a logistic function to model a binary response variable. We use feelings_life as the response variable which counts from 0 to 10. Different number represents different levels of satisfaction. If we change our response variable to the binary form 0 and 1, it can only show whether a person is happy or not instead of showing different levels of satisfaction. Therefore, we are not willing to change the response variable to a binary form so that we can not use the logistics regression model.

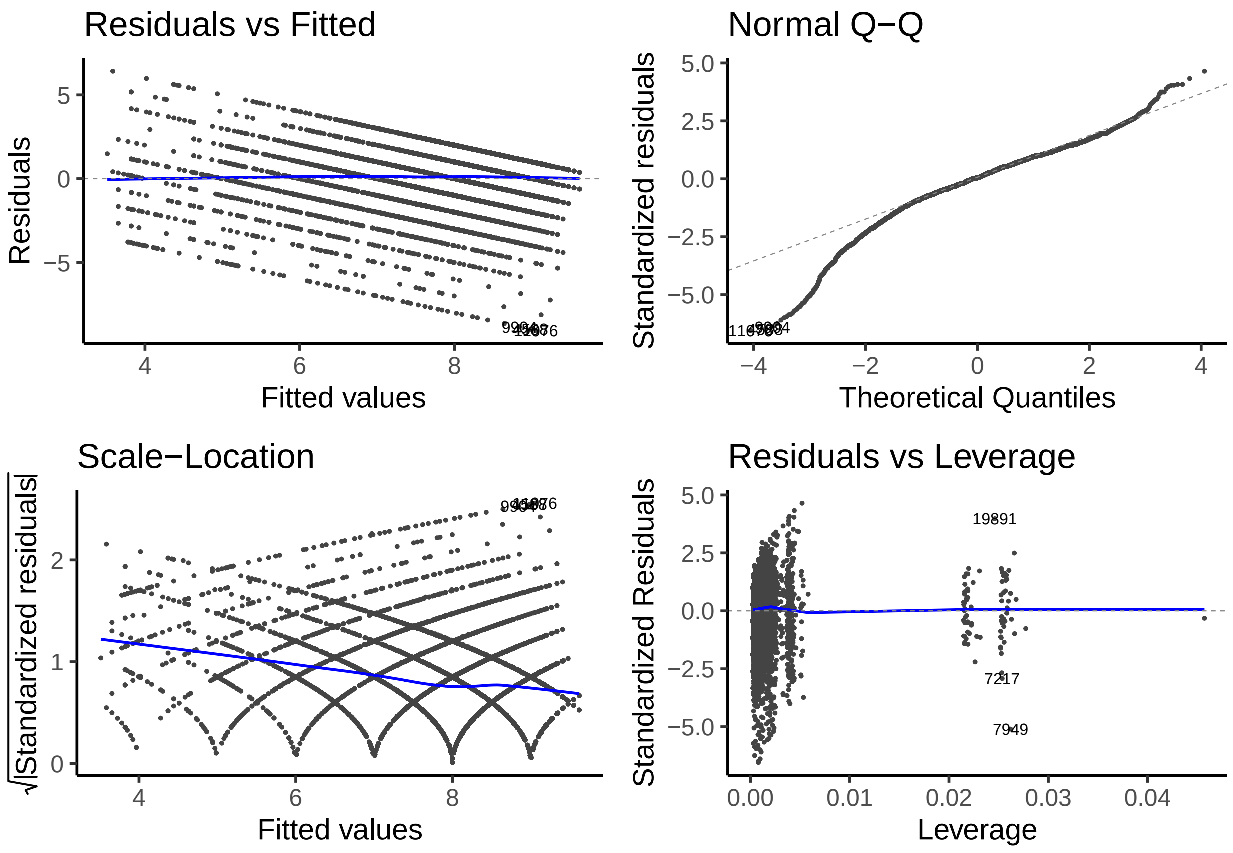
Our Linear Model:

$$\begin{aligned}
\hat{feelings\_life} = {} & 7.050402 + 0.353663 marital\_statusLivingcommon-law + 0.519051 marital\_statusMarried \\
& - 0.267279 marital\_statusSeparated + 0.042546 marital\_statusSingle, nevermarried \\
& + 0.238293 marital\_statusWidowed + 0.070904 total\_children \\
& + 0.831200 self\_rated\_healthExcellent + 0.067463 self\_rated\_healthFair \\
& + 0.421057 self\_rated\_healthGood \\
& - 0.740315 self\_rated\_healthPoor + 0.630628 self\_rated\_healthVerygood \\
& + 0.719563 self\_rated\_mental\_healthExcellent \\
& - 1.248459 self\_rated\_mental\_healthFair \\
& - 0.220055 self\_rated\_mental\_healthGood - 2.529230 self\_rated\_mental\_healthPoor \\
& + 0.273097 \beta_{16} self\_rated\_mental\_healthVerygood
\end{aligned}$$

Model Diagnostics

Below is the linear model diagnostics plots (Graph 6)

Graph 6 - Linear model diagnostics plot



Residuals vs Fitted

We can find out that there is a pattern in this residuals and fitted plot. Therefore, the shape of the pattern provides information on the function of x that is missing from the model.

Normal QQ plot

The normal QQ plot shows if residuals are normally distributed. We have known that if the residuals are lined well on the straight dashed line, the model we can fit is pretty good. From the normal QQ plot we made, we can find out that the residuals which are on the bottom left of this plot are not on the straight dashed line. However, most of the residuals are lined well on the straight line so that the model we would like to fit is good.

Scale-Location (Spread-Location plot)

The scale-location plot is also called spread-location plot. This plot can show if residuals are spread equally along the ranges of predictors. If we can find out a horizontal line with equally and randomly spread points, the model we can fit is pretty good. Besides, this plot can also assist us to check the assumption of equal variance. From the scale-location plot we made, we can find out that the residuals appear randomly spread so that the model we would like to fit is good.

Residuals vs Leverage

This kind of plot can help us to find the influential cases. We have known that not all outliers are influential in the linear regression analysis. Even though they have extreme values, they might not be important to determine a regression line which means that the results (model we would like to fit) would not be much different if we include or exclude them from analysis. However, some outliers can have an influence on the results. If there are some outliers at the upper right corner or at the lower right corner, the results can be influenced. However, in the residuals vs leverage plot we made, we did not find any points which are at the upper right corner of at the lower right corner so that the model we would like to fit is good.

Results

Model Summary

Here is the summary table for our model. (Summary table 1)

Summary table 1

```
##
## Call:
## lm(formula = feelings_life ~ marital_status + total_children +
##     self_rated_health + self_rated_mental_health, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0522 -0.7469  0.0804  0.9385  6.4155
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          7.050402   0.297351  23.711  < 2e-16 ***
## marital_statusLiving common-law      0.353663   0.045469   7.778 7.72e-15 ***
## marital_statusMarried                0.519051   0.036346  14.281  < 2e-16 ***
## marital_statusSeparated             -0.267279   0.064580  -4.139 3.51e-05 ***
## marital_statusSingle, never married  0.042546   0.040972   1.038 0.299090
## marital_statusWidowed                0.238293   0.046513   5.123 3.03e-07 ***
## total_children                       0.070904   0.007599   9.330  < 2e-16 ***
## self_rated_healthExcellent           0.831200   0.203748   4.080 4.53e-05 ***
## self_rated_healthFair                0.067463   0.204548   0.330 0.741545
## self_rated_healthGood                0.421057   0.202932   2.075 0.038011 *
## self_rated_healthPoor               -0.740315   0.208395  -3.552 0.000383 ***
## self_rated_healthVery good           0.630628   0.203061   3.106 0.001902 **
## self_rated_mental_healthExcellent    0.719563   0.220242   3.267 0.001088 **
## self_rated_mental_healthFair        -1.248459   0.222675  -5.607 2.09e-08 ***
## self_rated_mental_healthGood        -0.220055   0.219865  -1.001 0.316905
## self_rated_mental_healthPoor        -2.529230   0.233192 -10.846  < 2e-16 ***
## self_rated_mental_healthVery good    0.273097   0.219976   1.241 0.214441
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.385 on 20266 degrees of freedom
## Multiple R-squared:  0.2916, Adjusted R-squared:  0.2911
## F-statistic: 521.5 on 16 and 20266 DF,  p-value: < 2.2e-16
```

From the multiple linear regression model, we fit a linear regression model from feelings_life, marital_status, total_children,self_rated_health, and self_rated_mental_health:
we estimates this:

$$\begin{aligned}
\hat{feelings\_life} = \beta_0 &+ \beta_1 marital\_statusLivingcommon - law + \beta_2 marital\_statusMarried \\
&+ \beta_3 marital\_statusSeparated + \beta_4 marital\_statusSingle, nevermarried \\
&+ \beta_5 marital\_statusWidowed + \beta_6 total\_children \\
&+ \beta_7 self\_rated\_healthExcellent + \beta_8 self\_rated\_healthFair \\
&+ \beta_9 self\_rated\_healthGood \\
&+ \beta_{10} self\_rated\_healthPoor + \beta_{11} self\_rated\_healthVerygood \\
&+ \beta_{12} self\_rated\_mental\_healthExcellent \\
&+ \beta_{13} self\_rated\_mental\_healthFair \\
&+ \beta_{14} self\_rated\_mental\_healthGood + \beta_{15} self\_rated\_mental\_healthPoor \\
&+ \beta_{16} self\_rated\_mental\_healthVerygood
\end{aligned}$$

$$\begin{aligned}
\hat{feelings\_life} = 7.050402 &+ 0.353663 marital\_statusLivingcommon - law + 0.519051 marital\_statusMarried \\
&- 0.267279 marital\_statusSeparated + 0.042546 marital\_statusSingle, nevermarried \\
&+ 0.238293 marital\_statusWidowed + 0.070904 total\_children \\
&+ 0.831200 self\_rated\_healthExcellent + 0.067463 self\_rated\_healthFair \\
&+ 0.421057 self\_rated\_healthGood \\
&- 0.740315 self\_rated\_healthPoor + 0.630628 self\_rated\_healthVerygood \\
&+ 0.719563 self\_rated\_mental\_healthExcellent \\
&- 1.248459 self\_rated\_mental\_healthFair \\
&- 0.220055 self\_rated\_mental\_healthGood - 2.529230 self\_rated\_mental\_healthPoor \\
&+ 0.273097 \beta_{16} self\_rated\_mental\_healthVerygood
\end{aligned}$$

From the above regression line, we can interpret as if given other predictors holds constant, when marital_statusLiving common-law increase 1 unit, on average, feelings_life will increases by 0.353663 unit. This concept apply to other predictors in the same way, for example, given other predictors holds constant, marital_statusSeparated decrease 1 unit, on average, feelings_life will decrease by 0.267279 unit.

Hypothesis tests

In order to test whenever there is a significant linear relationship between feelings_life, marital_status, total_children,self_rated_health, and self_rated_mental_health. We need to do hypothesis testing for the estimates of the regression parameters.

For each slope estimates β_n:

$H_0 : \beta_n = 0$
$H_a : \beta_n \neq 0$

The null hypothesis states that the β_n is equal to zero, while the alternative hypothesis states that the β_n is not equal to zero.
In this case, we use a benchmark significance level of 5%

See the whole hypothesis test in Appendix

Here is our result of the test:

1. The significant features are: marital_statusLiving common-law, marital_statusMarried, marital_statusSeparated, marital_statusWidowed, total_children, self_rated_healthExcellent, self_rated_healthGood, self_rated_healthPoor, self_rated_healthVery good, self_rated_mental_healthExcellent, self_rated_mental_healthFair, and self_rated_mental_healthPoor correlates with feelings_life
2. The insignifiant features are: marital_statusSingle, never married, self_rated_healthFair, self_rated_mental_healthGood

Standard error

In addition, the standard error of the estimate is a measure of the accuracy of predictions.As we can see from the graph, the standard error of the estimates of all the parameter are relatively small. When the standard error is small, the sample data is said to be more representative of the true mean. Thus, we also can conclude that out predictions support the true relation of feelings_life, marital_status, total_children,self_rated_health, and self_rated_mental_health.

Residual standard error

Last but no least, the Residual standard error is the average distance that the response will deviate from the true regression line. In this case, we have residual standard error at 1.385 on 20266 degrees of freedom which tells us that average distance of the data points from the fitted line is about 1.385%. This a relatively small value; hence, we have evidence to suggest the linear model has predictive ability.

Discussion

Model intepretation

In term of hypothesis testing of the estimates of parameters in linear regression model, we can conclude that:

Marital status

We have evidence to support that feelings_life have correlations with marital_statusLiving common-law, marital_statusMarried, marital_statusSeparated, and marital_statusWidowed; while, we have no evidence to conclude that marital_statusSingle, never married has a relationship with feelings_life.
In general, marital status does affect people's feeling of life, among these statuses, living common-law and married status create a positive influence on feeling_life; however, separated status affect people's feeling about life negatively.
We can see that people who live common-law and married which are in a relatively stable conjugal relationship have a better feeling of life than divorced people on average. Besides, widower also has a positive effect on their feeling of life.

Total children

From the result of the hypothesis test of total children, we can conclude that we have strong evidence to support that there is a correlation between feelings_life and total children. In regression line, As total children increase one unit, on average, feelings_life has a slight increase by 0.070904 units.

In real life, the first child can increase a parent's happiness the most that parent of one child described higher life satisfaction than those without any children. However, the second child will not increase a parent's feeling of life as much as the first child. As a family have more children, they will have more economic pressures, more quarrel between family members; however, larger families may celebrate more meaning in their lives. Overall, we conclude that the happiness of parents increases as more children they have.

Self rated health and Self rated mental health

We have evidence to support that feelings_life have correlations with self_rated_healthExcellent, self_rated_healthGood, self_rated_healthPoor, self_rated_healthVery good; while we don't enough evidence to support this is a correlation with self_rated_healthFair.
For Self rated mental health, we also have strong evidence that self_rated_mental_healthExcellent, self_rated_mental_healthFair, and self_rated_mental_healthPoor correlates with feelings_life; however, there is no evidence for us to support self_rated_mental_healthGood correlates with feelings_life.
In the regression model, When people rate their health as excellent, their feeling of life increases the most as 0.831200 units, while those rate health as very good and good increase 0.630628 and 0.421057 respectively. However, for those who rate their health as poor, their feeling of life decreases 0.740315 units.
For mental health, people rate their mental health as excellent, their feeling of life increases the most as 0.719563 unit. However, when they rate their mental health as fair and poor, their feeling of life decreases 1.248459 and 2.529230 respectively.
As we can see, healthier or more metal healthier one person perceives, better life one person can feel. Since health is closely correlated with quality of life, bad physical and mental health unusually cause serious physical and mental diseases which often lead to unemployment, expensive health care and medical treatment and relatively short life span. The high cost of health care and medical treatment cause financial problems especially for those who are unable to work due to bad health or mental condition. These consequences can severely impact the quality of people's life. Thus, keeping a good physical and mental health condition is important for people to improve their feelings of life.

The source of the dataset and the bias

The General Social Survey (GSS) program conducted telephone surveys across the ten provinces. This survey mainly collected the family and living conditions of people aged 15 and above in Canada in order to form this GSS dataset.

The bias: There exists a response bias. Some people were not willing to answer all of the questions and some might not have a phone to answer the call so that the data contains a number of unavailable data (NA) which can have an influence on the accuracy of our analysis and conclusions.

A discussion of the questionnaire of this survey

The advantages of the questionnaire are obvious. It involves many aspects of living conditions such as marital status, health condition, and income. Therefore, it provides an overview about the basic living conditions of people aged 15 and over in Canada.

However, the questionnaire also exists some drawbacks. For instance, the question "Using a scale of 0 to 10 means"Very dissatisfied" and 10 means "Very satisfied", how do you feel about your life as a whole right now?". For the option of this question, 0 means very dissatisfied, 5 means generally satisfied and 10 means very satisfied. However, some people who were not satisfied with their life might select 0, 1 or 2 so that this choice makes the resulting data become less accurate. Besides, some questions include privacy issues so that a number of respondents are not willing to disclose their personal information or provide inaccurate data instead. Either one of these situations may influence the accuracy of our analysis and conclusion.

Conclusions

The goal of our analysis is to explore how family (marital status and total children) and health (self-rated health and self-rated mental health) can affect respondents' feeling of life in the GSS dataset as feelings_life is a good indicator to reflect citizens' standard of living.

In our report, we discussed some key features, strengths and weaknesses of the data we use firstly. Then, we select two main features (family and health) to fit a multiple linear regression model to estimate the response variable (feelings_life) we chose. In this part, we also use model diagnostics to show that whether we can fit a good model or not. Besides, we use feature and model selection in the appendix. Next, we present our results and explain them. Eventually, we discuss the survey, data and the model and identify the strengths and weaknesses.

Overall, we find out that a more stable conjugal relationship will increase people's feeling of life; however, we find out that widower also tends to have greater happiness of life which is something surprise us in the result. Besides, people who are healthier in physical and the mental condition tends to have a better quality of life which also reflects as a better feeling of life.

Weaknesses

The weakness of data

We can find out that some people did not response the question since some data in the dataset are unavailable (NA). Besides, some people might answer the questions casually or incorrectly which can cause an error. Therefore, it can have an influence on the accuracy of our analysis so that the weakness of the data is that non-response or response error could exist.

The weakness of our analysis

The model we fit in this report is multiple linear regression model which use four independent variables to estimate one response variable called feelings_life. However, we can hardly measure linearity since we just use model selection (See Appendix for more about how we select features) to select the most representative features which can influence feelings_life instead of measuring the linearity. Therefore, the weakness of our analysis is that the linear trend is hard to be estimated.

Next Steps

More Analysis:

1. We can try more models for a better demonstration of the relationships. For example, Generalized Additive Model may be helpful for exploring non-linear relationships. The function term in this model may have a positive effect in our model accuracy.
2. We can also add more features to our model and penalize overfitting by methods such as Ridge regression and LASSO regression. This can improve our feature selection and at the same time identifying the most significant features.
3. We can train models and do some predictions, which is useful not only in knowing more about the performance about our model, but also in making predictions when we get more data and assess their feelings of lives. In order to train a more accurate model, machine learning techniques such as neural network may be a good choice.

References

"Welcome to My.access – Please Choose How You Will Connect." My.access - University of Toronto Libraries Portal,
sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/ hsda?harcsda4+gss3. Hayes, Adam. "Reading Into Stratified Random Sampling." Investopedia, Investopedia, 16 Sept. 2020, www.investopedia.com/terms/stratified_rando

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 Wickham et al., (2019). Welcome to the tidyverse. Journal of Open

Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 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

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Appendix

Feature selection

First of all, we chose a response variable that we are interested in: feelings of life.

Next, we identified several parts of possible features:

1. Basic information: age, sex, region (make sure that the survey is evenly distributed between these categories, making the inference representative)
2. Family: marital_status, total_children
3. Living condition: own_rent, living_arrangement, hh_type, hh_size
4. Health: self_rated_health, self_rated_mental_health
5. religon: regilion_importance
6. Knowledge and education: language_knowledge, education
7. Financial status: income_family, income_respondent

We clean the data select these features above into a single data frame.

We then implemented backward BIC (Bayesian Information Criterion) to select the features. This method involves choosing the optimum model by deleting one most insignificant feature a time. BIC is a penalized cost function for the model. After several steps, when the value of BIC reaches its minimum, we get our optimum model.

Table 6: Table 7 - BIC

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	19793	36771.38	12897.07
- living_arrangement	11	95.0117343	19804	36866.40	12839.45
- income_respondent	5	7.9471800	19809	36874.34	12794.25
- hh_type	4	9.6277025	19813	36883.97	12759.85
- language_knowledge	4	35.4538856	19817	36919.42	12739.35
- hh_size	1	0.0579119	19818	36919.48	12729.48
- region	4	58.8041571	19822	36978.29	12721.50
- income_family	5	84.0528236	19827	37062.34	12717.11

From the table above (Table 7), We can see that seven features are deleted from the model. However, this model also has too many features for us to work with. We are looking for some features that are most influential to feelings of life.

We then turn to choose from the remaining model by assessing the p-values of the coefficients. We tend to choose features with smaller p-value, which means more likely to reject the null hypothesis of coefficients being zero. The summary table of our model after using BIC is below.

Table 7: Table 8 - p-values

names	pvalue
(Intercept)	< 2e-16
age	3.08e-07
as.factor(sex)Male	6.84e-08
pop_centerPrince Edward Island	0.020967
pop_centerRural areas and small population centres (non CMA/CA)	2.68e-05
marital_statusLiving common-law	< 2e-16

names	pvalue
marital_statusMarried	< 2e-16
marital_statusSeparated	5.32e-05
marital_statusSingle, never married	0.026131
marital_statusWidowed	0.061447
total_children	3.3e-06
own_rentOwned by you or a member of this household, even if it i…	0.957786
own_rentRented, even if no cash rent is paid	0.45578
self_rated_healthExcellent	2.83e-06
self_rated_healthFair	0.441573
self_rated_healthGood	0.008111
self_rated_healthPoor	0.002519
self_rated_healthVery good	0.00022
self_rated_mental_healthExcellent	2.83e-05
self_rated_mental_healthFair	1.87e-05
self_rated_mental_healthGood	0.928233
self_rated_mental_healthPoor	< 2e-16
self_rated_mental_healthVery good	0.024585
regilion_importanceNot at all important	0.064259
regilion_importanceNot very important	0.050481
regilion_importanceSomewhat important	0.080718
regilion_importanceVery important	0.834614
educationCollege, CEGEP or other non-university certificate or di…	0.091784
educationHigh school diploma or a high school equivalency certificate	0.000143
educationLess than high school diploma or its equivalent	< 2e-16
educationTrade certificate or diploma	0.020429
educationUniversity certificate or diploma below the bachelor's level	0.124352
educationUniversity certificate, diploma or degree above the bach…	0.694212

It can be seen from the above table (Table 8) that the p-values are smallest in terms `marital_status, total_children, self_rated_health, self_rated_mental_health`. Thus we choose these features for our final linear model.

Hypothesis tests

Hypothesis Testing of intercept estimate.
$H_0 : \beta_0 = 0$
$H_a : \beta_0 \neq 0$
The null hypothesis states that the β_0 is equal to zero, while the alternative hypothesis states that the β_0 is not equal to zero.
In this case, we use a benchmark significance level of 5%;thus,the p_value of the intercept estimate is 2e-16 which is extremely smaller than 0.05.
As a result, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and intercept estimate.

Hypothesis Testing of β_1 estimate.
$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$
The null hypothesis states that the β_1 is equal to zero, while the alternative hypothesis states that the β_1 is not equal to zero.
Since the p_value of β_1 is 7.72e-15 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and marital_statusLiving

common-law.

Hypothesis Testing of β_2 estimate.
$H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$
Since the p_value of β_2 is 2e-16 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and marital_statusMarried.

Hypothesis Testing of β_3 estimate.
$H_0 : \beta_3 = 0$
$H_a : \beta_3 \neq 0$
Since the p_value of β_3 is 3.51e-05 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and marital_statusSeparated.

However, we get different result from hypothesis testing of β_4 estimate:
$H_0 : \beta_4 = 0$
$H_a : \beta_4 \neq 0$
Since the p_value of β_4 is 0.299090 which is larger than 0.05, we don't reject H_0 which means that indicates that there is no evidence for us to support the correlation of feelings_life and marital_statusSingle, never married.

Hypothesis Testing of β_5 estimate.
$H_0 : \beta_5 = 0$
$H_a : \beta_5 \neq 0$
Since the p_value of β_5 is 3.03e-07 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and marital_statusWidowed.

Hypothesis Testing of β_6 estimate.
$H_0 : \beta_6 = 0$
$H_a : \beta_6 \neq 0$
Since the p_value of β_6 is2e-16 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and total_children.

Hypothesis Testing of β_7 estimate.
$H_0 : \beta_7 = 0$
$H_a : \beta_7 \neq 0$
Since the p_value of β_7 is 4.53e-05 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and self_rated_healthExcellent.

Hypothesis Testing of β_8 estimate.
$H_0 : \beta_8 = 0$
$H_a : \beta_8 \neq 0$
Since the p_value of β_8 is 0.741545 which is larger than 0.05, we don't reject H_0 which means that indicates that there is no evidence for us to support the correlation of feelings_life and self_rated_healthFair.

Hypothesis Testing of β_9 estimate.
$H_0 : \beta_9 = 0$
$H_a : \beta_9 \neq 0$
Since the p_value of β_9 is 0.038011 which is smaller than 0.05, we reject H_0 in favor of the alternative

hypothesis which indicates that there is correlation of feelings_life and self_rated_healthGood.

Hypothesis Testing of β_{10} estimate.
$H_0 : \beta_{10} = 0$
$H_a : \beta_{10} \neq 0$
Since the p_value of β_{10} is 0.000383 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and self_rated_healthPoor.

Hypothesis Testing of β_{11} estimate.
$H_0 : \beta_{11} = 0$
$H_a : \beta_{11} \neq 0$
Since the p_value of β_{11} is 0.001902 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and self_rated_healthVery good.

Hypothesis Testing of β_{12} estimate.
$H_0 : \beta_{12} = 0$
$H_a : \beta_{12} \neq 0$
Since the p_value of β_{12} is 0.001088 which is smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is correlation of feelings_life and self_rated_mental_healthExcellent.

Hypothesis Testing of β_{13} estimate.
$H_0 : \beta_{13} = 0$
$H_a : \beta_{13} \neq 0$
Since the p_value of β_{13} is 2.09e-08 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and self_rated_mental_healthFair.

Hypothesis Testing of β_{14} estimate.
$H_0 : \beta_{14} = 0$
$H_a : \beta_{14} \neq 0$
Since the p_value of β_{14} is 0.316905 which is larger than 0.05, we don't reject H_0 which means that indicates that there is no evidence for us to support the correlation of feelings_life and self_rated_mental_healthGood.

Hypothesis Testing of β_{15} estimate.
$H_0 : \beta_{15} = 0$
$H_a : \beta_{15} \neq 0$
Since the p_value of β_{15} is 2e-16 which is extremely smaller than 0.05, we reject H_0 in favor of the alternative hypothesis which indicates that there is significant correlation of feelings_life and self_rated_mental_healthPoor.

Hypothesis Testing of β_{16} estimate.
$H_0 : \beta_{16} = 0$
$H_a : \beta_{16} \neq 0$
Since the p_value of β_{16} is 0.214441 which is larger than 0.05, we don't reject H_0 which means that indicates that there is no evidence for us to support the correlation of feelings_life and self_rated_mental_healthVery good.