What Makes Canadian Rate Their Feelings of Life

STA304 Problem Set 2

Yiwen Feng 1004890932, Yutong Yuan 1004725103,

Sirui Xu 1004963603, Man Fei 1002129984

Code and data supporting this analysis is available at:

"https://github.com/otisfeng/sta304-ps2-group24"

Abstract

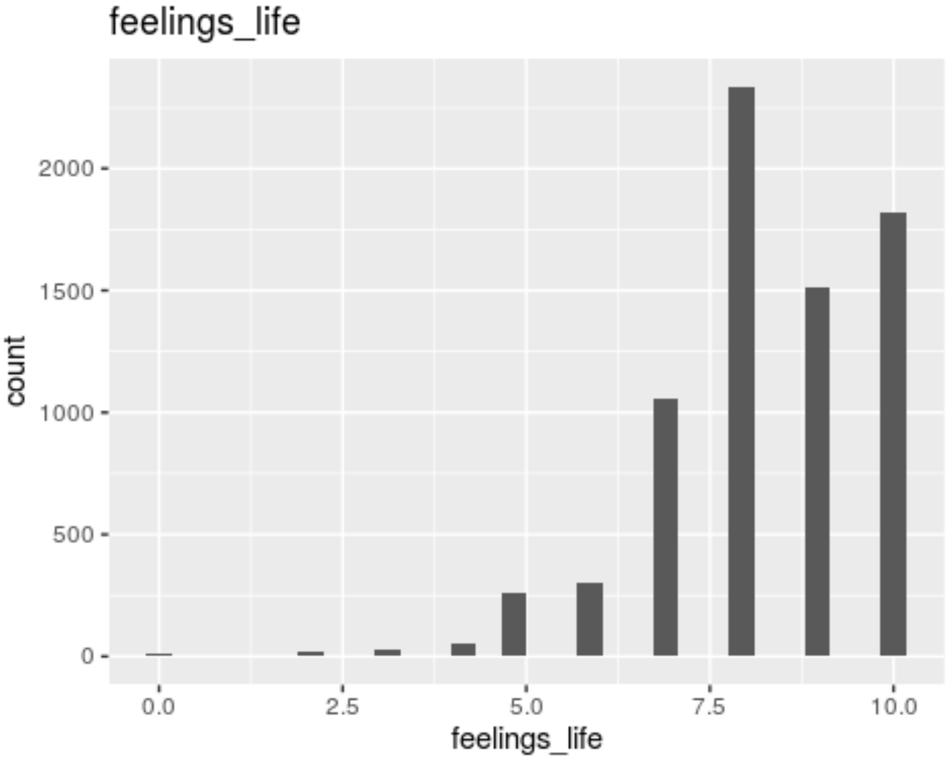
In order to observe the changes in the living conditions and satisfaction of Canadians overtime, our group use linear model to find out the influence of various variables to the respondent's feelings about life as a whole using GSS(General Social Survey) on the Family data. During the progressing of building the model, we figured out the most important aspect of respondents' feelings is about their health and mental health. Overall, we conclude that all the factors in the research have linear relationship.

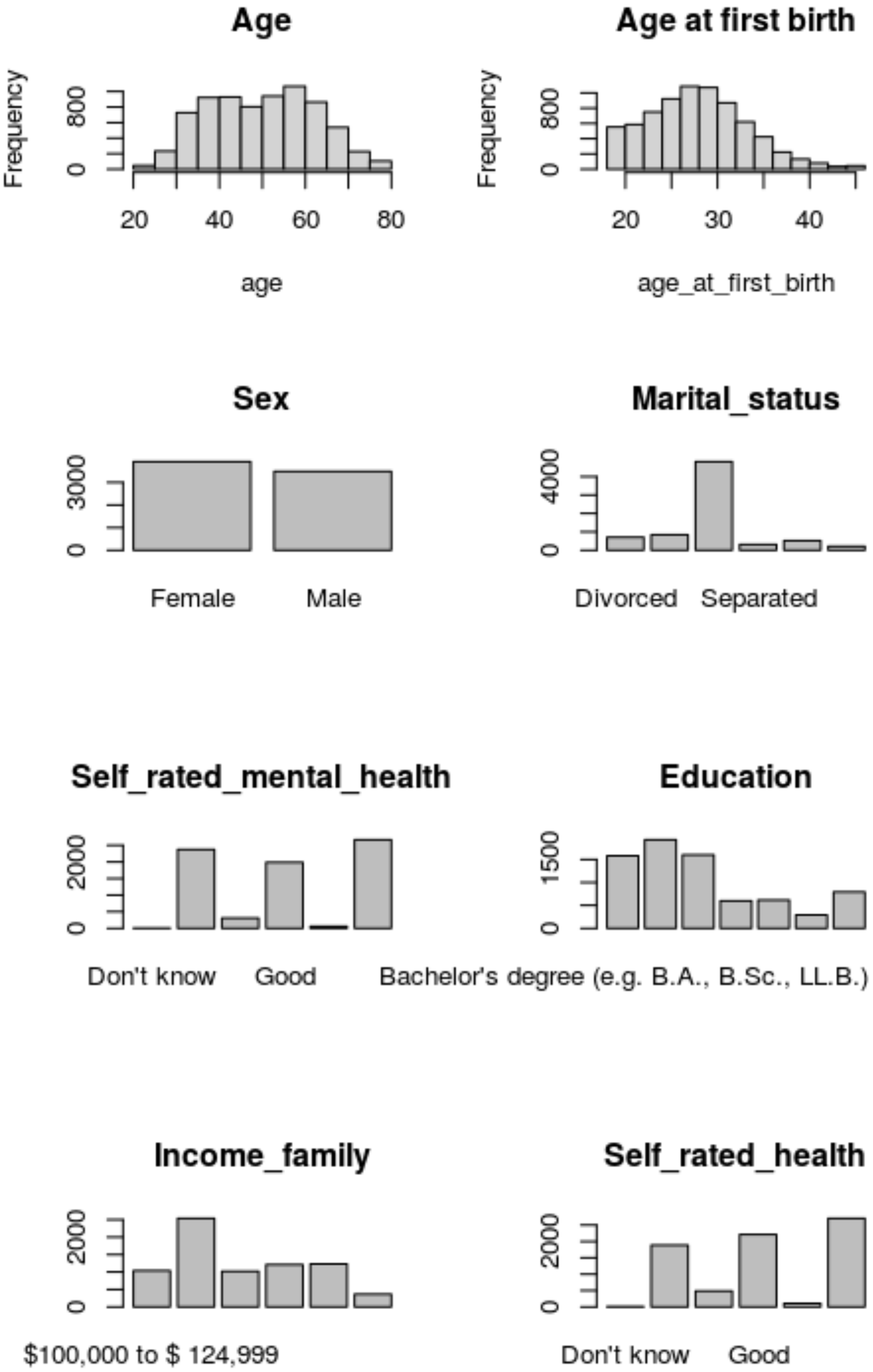
Introduction

Feelings about life is probably what matters the most for a human being. In this report, we are wondering if the happiness (Feelings_life) for Canadian relates to their financial situation, education level, family situation(marriage, children), mental & physical health, and sex.

Therefore, we collected the dataset from the 2017 Canadian General Social Survey (Cycle 31, Family) to study if there is a multiple linear relationship between Canadians' Age, Feelings_life, age_at_first_birth, sex, marital_status, education, income_family, self_rated_health, and self_rated_mental_health.

We found that almost all the variables above have linear relationship with feelings life. But the excellence of mental health is what influence the most. It is easy to interpret why is it important since a good mental health is what someone's feelings about life based on. Within this report, We will talk about the survey method, the model we build, and share results and the possible causes of the linear relationship we obtained. Also, we will state the limitation and next step if we would like to investigate more about happiness for Canadian.




Age
Frequency
0
800
20
40
60
80
age
Age at first birth
Frequency
0
800
20
30
40
age_at_first_birth
Sex
0
3000
Female
Male
Marital_status
0
4000
Divorced
Separated
Self_rated_mental_health
0
2000
Don't know
Good
Education
0
1500
Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
Income_family
0
2000
$100,000 to $ 124,999
Self_rated_health
0
2000
Don't know
Good

Data

The source of the data.

We obtain a dataset from the 2017 General Social Survey (GSS) on the Family collected by Statistics Canada.

-The methodology and approach that is used to collect and process the data.
To carry out the sampling, GSS used stratified sampling. People were divided into 27 strata which include ten provinces strata plus ten more non-CMA(Census Metropolitan Areas) area strata. This is a good sampling method for the large target population, especially based on the entire country. In this case, we can control the regional differences and do a simple random sample in each stratum.

-The population, the frame, and the sample.
The target population for 2017 GSS is all persons who are 15 years of age and older in Canada excluding the residents of the Yukon, Northwest Territories, and Nunavut, and the Full-time residents of institutions.
The survey frame included two-component, a list of telephone numbers in use available to Statistics Canada from various sources and The Address Register (AR) (list of all dwellings within the ten provinces). The central role of GSS is family, so all telephone numbers associated with the same valid address can be grouped by Address Register as a family.
The target sample size for 2017 GSS was 20,000 while the actual number of respondents was 20,602. The dataset was collected by computer-assisted telephone interviews (CATI) which is made from approximately 9:00 a.m. to 9:30 p.m. Mondays to Fridays, 10:00 a.m. to 5:00 p.m. on Saturdays and 1:00 p.m. to 9:00 p.m. on Sundays with the sampling population.
Also, for people who refused or not being convenient to answer the interviewer's call, reconnection and appointment were made for them to make sure a good coverage of all households with telephone numbers. Finally, the overall response rate for 2017 GSS was 52.4%.
the non-response are clearly marked as "Valid skip", "don't know", "Refusal" and "Not stated" in the survey.

strengths and weaknesses

Strengths: It was investigate a very large sample frame, so the data the survey gathered should be quiet representative it self. Also, the Non-Responses has been clearly indicated in the survey.
Weakness: Most of the questions that the survey asked are quiet private, so there could be a lot of people who lied about it. (Response- Error)

Variables

In our group assignment report, the elements we took are Age, Feelings_life, age_at_first_birth, sex, marital_status, education, income_family, self_rated_health, and self_rated_mental_health.

1. "age" was described as "the age of respondent at the time of the survey interview" in the questionnaire concept, with the universe of all respondents but was capped at 80 years. The histogram shows two modes, one at age 40 and one at age 60; which means most of the people we surveied are at their 40s and 60s

2. "feelings_life" means feelings about life as a whole. The question text is "Using a scale of 0 to 10 where 0 means 'Very dissatisfied' and 10 means 'Very satisfied', how do you feel about your life as a whole right now?". The scale is a suitable way to measure satisfaction in the questionnaire, and the collective result will be numeric which can be displayed as the visual model when we analyzing the survey result. The histogram is very left skewed, most of people rate 8 to 10 (very satisfied) about their lives, only a few people rated under 5.

3. "age_at_first_birth" is the age of respondents when they gave birth to their first child. And the answer categories were people whose ages are above 15 years. The questionnaire concept based on those three elements was effective and meaningful. The questions are concise and clear enough to understand, and the answers collected are numeric which is much easier for researchers to make classification and analysis on Canadian social trends. The histogram is very right skewed with most of people's age at first birth were around 28.

4. ”sex”, a categorical variable. The code for answer “Female” is 2 while the code for “Male” is 1, the variable is categorical. The collective result will be numeric which can be displayed as the visual model when we analyzing the survey result. The bar plot represents the count of different sex has been included above. We can learn that their are almost equal amount for each sex.

5. “marital_status”, categorical variable, the answers categories are “Married”, “Living common-law”, “Widowed”, “Separated”, “Divorced”, “Single, never married”, these answers are corresponded with codes “01” to “06”. Also, “Valid skip” = code “96”, ”don’t know” = code “97”, “Refusal” = code “98”, and “Not stated” code “99”. The collective results are numeric and easy to be analyzed. The bar plot represents the count of different marital_status has been included above. We can learn that most of people are married from the barplot.

6. “Education”, categorical variable.The question text is “What is the highest certificate, diploma or degree that you have completed?”. The answers are varies from “Less than high school diploma” to “University certificate, diploma or degree”, with corresponding codes from “01” to “06”. Also, “Valid skip” = code “96”, “don’t know” = code “97”, “Refusal” = code “98”, and “Not stated” code “99”.The collective results are numeric and easy to be analyzed. The bar plot represents the count of different education level has been included above.

7. ”income_family” : categorical variable, answers are varies from “Less than $25,000” to “$125,000 and more”, with corresponded codes from “01” to “06”. Also, “Valid skip” = code “96”, “don’t know” = code “97”, “Refusal” = code “98”, and “Not stated” code “99”. The collective results are numeric and easy to be analyzed. The bar plot tells us that there are a lot of people earns $125,000 and more, the rest of the income level are distributed similarly.

8. “self_rated_health”: categorical variable, questionnaire questions is “In general, would you say your health is⋯?”. The answers are varies from “Excellent” to “ Poor”, with codes from “1” to “5”. Also, “Valid skip” = code “6”, “don’t know” = code “7”, “Refusal” = code “8”, and “Not stated” code “9”. The scale is a suitable way to measure health in the questionnaire, and the collective result will be numeric which can be displayed as the visual model when we analyzing the survey result. The bar plot tells us that most of people feel very good, good, and excellent about their health. Only a few people do not know or rated poor for themselves.

9. “self_rated_mental_health”: categorical variable, questionnaire questions is “In general, would you say your mental health is⋯?”. The answers are exactly follow the same patterns as the previous variable “self_rated_health”. The bar plot tells us that most of people feel very good, good, and excellent about their mental health. Only a few people do not know or rated poor for themselves.

Model

```
##
## Call:
## lm(formula = feelings_life ~ self_rated_health + self_rated_mental_health,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0602 -0.6828  0.1249  0.9398  6.1175
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        7.0253     0.6142  11.437  < 2e-16 ***
## self_rated_healthExcellent         0.7209     0.4108   1.755  0.07933 .
## self_rated_healthFair              0.1569     0.4136   0.379  0.70439
## self_rated_healthGood              0.3979     0.4102   0.970  0.33208
## self_rated_healthPoor             -0.2636     0.4294  -0.614  0.53932
## self_rated_healthVery good         0.5358     0.4100   1.307  0.19129
## self_rated_mental_healthExcellent  1.3139     0.4591   2.862  0.00422 **
## self_rated_mental_healthFair      -0.7008     0.4640  -1.510  0.13099
## self_rated_mental_healthGood       0.2596     0.4586   0.566  0.57137
## self_rated_mental_healthPoor      -2.8792     0.4889  -5.889 4.04e-09 ***
## self_rated_mental_healthVery good  0.8710     0.4587   1.899  0.05763 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.294 on 7398 degrees of freedom
## Multiple R-squared:  0.2296, Adjusted R-squared:  0.2286
## F-statistic: 220.5 on 10 and 7398 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = feelings_life ~ age_at_first_birth + age + sex +
##     marital_status + education + income_family + self_rated_health +
##     self_rated_mental_health, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5060 -0.7202  0.0801  0.8793  6.2347
##
## Coefficients:
##                                                                   Estimate
## (Intercept)                                                       6.978078
## age_at_first_birth                                               -0.015671
## age                                                               0.005035
## sexMale                                                          -0.110156
## marital_statusLiving common-law                                   0.340159
## marital_statusMarried                                             0.436738
## marital_statusSeparated                                          -0.180917
## marital_statusSingle, never married                              -0.106822
## marital_statusWidowed                                            -0.266238
## educationCollege, CEGEP or other non-university certificate or di...  0.060929
## educationHigh school diploma or a high school equivalency certificate  0.102985
## educationLess than high school diploma or its equivalent          0.330906
## educationTrade certificate or diploma                             0.159688
## educationUniversity certificate or diploma below the bachelor's level  0.088087
## educationUniversity certificate, diploma or degree above the bach...  -0.030221
## income_family$125,000 and more                                        0.044656
## income_family$25,000 to $49,999                                      -0.210183
## income_family$50,000 to $74,999                                      -0.023073
## income_family$75,000 to $99,999                                      -0.093609
## income_familyLess than $25,000                                       -0.252708
## self_rated_healthExcellent                                            0.899280
## self_rated_healthFair                                                 0.288198
## self_rated_healthGood                                                 0.558666
## self_rated_healthPoor                                                -0.070201
## self_rated_healthVery good                                            0.695305
## self_rated_mental_healthExcellent                                     1.084089
## self_rated_mental_healthFair                                         -0.835097
## self_rated_mental_healthGood                                          0.054305
## self_rated_mental_healthPoor                                         -2.902332
## self_rated_mental_healthVery good                                     0.658690
##                                                                     Std. Error
## (Intercept)                                                           0.615133
## age_at_first_birth                                                    0.002983
## age                                                                   0.001296
## sexMale                                                               0.031401
## marital_statusLiving common-law                                       0.068761
## marital_statusMarried                                                 0.055537
## marital_statusSeparated                                               0.086023
## marital_statusSingle, never married                                   0.075057
## marital_statusWidowed                                                 0.099373
## educationCollege, CEGEP or other non-university certificate or di...   0.043918
## educationHigh school diploma or a high school equivalency certificate   0.047448
## educationLess than high school diploma or its equivalent                0.065528
## educationTrade certificate or diploma                                   0.061754
## educationUniversity certificate or diploma below the bachelor's level   0.080982
## educationUniversity certificate, diploma or degree above the bach...    0.055220
## income_family$125,000 and more                                        0.047064
## income_family$25,000 to $49,999                                       0.059613
## income_family$50,000 to $74,999                                       0.054591
## income_family$75,000 to $99,999                                       0.053508
## income_familyLess than $25,000                                        0.081945
## self_rated_healthExcellent                                            0.402247
## self_rated_healthFair                                                 0.404600
## self_rated_healthGood                                                 0.401443
## self_rated_healthPoor                                                 0.420073
## self_rated_healthVery good                                            0.401347
## self_rated_mental_healthExcellent                                     0.449095
## self_rated_mental_healthFair                                          0.453856
## self_rated_mental_healthGood                                          0.448627
## self_rated_mental_healthPoor                                          0.477782
## self_rated_mental_healthVery good                                     0.448737
##                                                                     t value
## (Intercept)                                                          11.344
## age_at_first_birth                                                   -5.254
## age                                                                   3.884
## sexMale                                                              -3.508
## marital_statusLiving common-law                                       4.947
## marital_statusMarried                                                 7.864
## marital_statusSeparated                                              -2.103
## marital_statusSingle, never married                                  -1.423
## marital_statusWidowed                                                -2.679
## educationCollege, CEGEP or other non-university certificate or di...    1.387
## educationHigh school diploma or a high school equivalency certificate   2.170
## educationLess than high school diploma or its equivalent                5.050
## educationTrade certificate or diploma                                   2.586
## educationUniversity certificate or diploma below the bachelor's level   1.088
## educationUniversity certificate, diploma or degree above the bach...   -0.547
## income_family$125,000 and more                                        0.949
## income_family$25,000 to $49,999                                      -3.526
## income_family$50,000 to $74,999                                      -0.423
## income_family$75,000 to $99,999                                      -1.749
## income_familyLess than $25,000                                       -3.084
## self_rated_healthExcellent                                            2.236
## self_rated_healthFair                                                 0.712
## self_rated_healthGood                                              1.392
## self_rated_healthPoor                                             -0.167
## self_rated_healthVery good                                         1.732
## self_rated_mental_healthExcellent                                  2.414
## self_rated_mental_healthFair                                      -1.840
## self_rated_mental_healthGood                                       0.121
## self_rated_mental_healthPoor                                      -6.075
## self_rated_mental_healthVery good                                  1.468
##                                                             Pr(>|t|)
## (Intercept)                                                  < 2e-16
## age_at_first_birth                                          1.53e-07
## age                                                         0.000104
## sexMale                                                     0.000454
## marital_statusLiving common-law                             7.70e-07
## marital_statusMarried                                       4.25e-15
## marital_statusSeparated                                     0.035488
## marital_statusSingle, never married                         0.154719
## marital_statusWidowed                                       0.007397
## educationCollege, CEGEP or other non-university certificate or di...  0.165381
## educationHigh school diploma or a high school equivalency certificate 0.030002
## educationLess than high school diploma or its equivalent              4.53e-07
## educationTrade certificate or diploma                                 0.009732
## educationUniversity certificate or diploma below the bachelor's level 0.276749
## educationUniversity certificate, diploma or degree above the bach...  0.584205
## income_family$125,000 and more                              0.342737
## income_family$25,000 to $49,999                             0.000425
## income_family$50,000 to $74,999                             0.672565
## income_family$75,000 to $99,999                             0.080257
## income_familyLess than $25,000                              0.002051
## self_rated_healthExcellent                                  0.025405
## self_rated_healthFair                                       0.476299
## self_rated_healthGood                                       0.164072
## self_rated_healthPoor                                       0.867284
## self_rated_healthVery good                                  0.083239
## self_rated_mental_healthExcellent                           0.015805
## self_rated_mental_healthFair                                0.065808
## self_rated_mental_healthGood                                0.903656
## self_rated_mental_healthPoor                                1.30e-09
## self_rated_mental_healthVery good                           0.142181
##
## (Intercept)                                                 ***
## age_at_first_birth                                          ***
## age                                                         ***
## sexMale                                                     ***
## marital_statusLiving common-law                             ***
## marital_statusMarried                                       ***
## marital_statusSeparated                                     *
## marital_statusSingle, never married
## marital_statusWidowed                                       **
## educationCollege, CEGEP or other non-university certificate or di...
## educationHigh school diploma or a high school equivalency certificate *
## educationLess than high school diploma or its equivalent              ***
## educationTrade certificate or diploma                                 **
## educationUniversity certificate or diploma below the bachelor's level
## educationUniversity certificate, diploma or degree above the bach...
## income_family$125,000 and more
## income_family$25,000 to $49,999                             ***
## income_family$50,000 to $74,999
## income_family$75,000 to $99,999                             .
## income_familyLess than $25,000                              **
## self_rated_healthExcellent                                  *
## self_rated_healthFair
## self_rated_healthGood
## self_rated_healthPoor
## self_rated_healthVery good                                  .
## self_rated_mental_healthExcellent                           *
## self_rated_mental_healthFair                                .
## self_rated_mental_healthGood
## self_rated_mental_healthPoor                                ***
## self_rated_mental_healthVery good
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.262 on 7379 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2659
## F-statistic: 93.54 on 29 and 7379 DF,  p-value: < 2.2e-16
```

We continue our analysis by building a multiple linear regression model to predict if there is a multiple linear relationship between Canadians' "feelings_life" and "age_at_first_birth", "age", "sex", "marital_status", "education", "income_family", "self_rated_health", "self_rated_mental_health". A multiple linear regression model is trying to model the relationship between two or more explanatory variables x and a response variable y by fitting a linear equation to observed data. In this report, we use seven independent variables x, and a dependent response variable y. So, our estimated multiple linear regression can be expressed as
y = Bata0 + Bata1x1 + Bata2x2 +Bata3x3 + Bata4x4 + Bata5x5 + Bata6x6 + Bata7x7 + Beta8x8 + residuals

The summary of the model generated is:

notation	variable name	coefficient	coefficient value	p-value
b_0	(Intercept)	*b0*	6.978078	< 2e-15
x1	age_at_first_birth	*b1*	-0.015671	1.53E-07
x2	age	*b2*	0.005035	0.000104
x3	sexMale	*b3*	-0.110156	0.000454
x4	marital_statusLiving common-law	*b4*	0.340159	7.70E-07
x5	marital_statusMarried	*b5*	0.436738	4.25E-15
x6	marital_statusSeparated	*b6*	-0.180917	0.035488
x7	marital_statusSingle, never married	*b7*	-0.106822	0.154719
x8	marital_statusWidowed	*b8*	-0.266238	0.007397
x9	educationCollege, CEGEP or other non-university certificate or diploma	*b9*	0.060929	0.165381
x10	educationHigh school diploma or a high school equivalency certificate	*b10*	0.102985	0.030002
x11	educationLess than high school diploma or its equivalent	*b11*	0.330906	4.53E-07
x12	educationTrade certificate or diploma	*b12*	0.159688	0.009732
x13	educationUniversity certificate or diploma below the bachelor's level	*b13*	0.088087	0.276749
x14	educationUniversity certificate, diploma or degree above the bachelor’s level	*b14*	-0.030221	0.584205
x15	income_family$125,000 and more	*b15*	0.044656	0.342737
x16	income_family$25,000 to $49,999	*b16*	-0.210183	0.000425
x17	income_family$50,000 to $74,999	*b17*	-0.023073	0.672565
x18	income_family$75,000 to $99,999	*b18*	-0.093609	0.080257
x19	income_familyLess than $25,000	*b19*	-0.252708	0.002051
x20	self_rated_healthExcellent	*b20*	0.899280	0.025405
x21	self_rated_healthFair	*b21*	0.288198	0.476299
x22	self_rated_healthGood	*b22*	0.558666	0.164072
x23	self_rated_healthPoor	*b23*	-0.070201	0.867284
x24	self_rated_healthVery good	*b24*	0.695305	0.083239
x25	self_rated_mental_healthExcellent	*b25*	1.084089	0.015805
x26	self_rated_mental_healthFair	*b26*	-0.835097	0.065808
x27	self_rated_mental_healthGood	*b27*	0.054305	0.903656
x28	self_rated_mental_healthPoor	*b28*	-2.902332	1.30E-09
x29	self_rated_mental_healthVery good	*b29*	0.658690	0.142181

Form 1:
table appendix: The table shows all the variable name from x1 to x29, the corresponding coefficient b and b value, and their p value. We get all these data from the multiple linear regression ran by R studio.

In this form, we can find that
from x4 to x8 are the dummy variables of “marital_status” with the baseline of “marital_status_Divorced”;
from x9 to x14 are the dummy variables of “education” with the baseline of “educationBachelor’s degree (e.g. B.A., B.Sc., LL.B.)”;
from x15 to x19 are the dummy variables of “income_family” with the baseline of “income_family$99,999 to $125,000”;
from x20 to x24 are the dummy variables of “self_rated_health”;
from x25 to x29 are the dummy variable of “self_rated_mental_health”.

Which then results in the following model:

$$\hat{y}_{(feelingsoflife)}$$
$$= b0 + b1x_{age_{a}t_{f}irst_{b}irth} + b2x_{age} + b3x_{sexMale} + b4x_{marital_{s}tatusLivingcommon-law} + \cdots$$
$$+ b29x_{self_{r}ated_{m}ental_{h}ealthVerygood}$$

Where the response variable y hat is our predicted respondents' feeling of life. On the table, we could find all the variable name from x1 to x29, the corresponding coefficient b and b value, and their p value. In our estimation model, the number of independent variable x was increased to twenty-nine, the reason is that some of our elements (eg. "marital_status", "education", "income_family", "self_rated_health", "self_rated_mental_health") are categorical variable, when we import these categorical variables into R studio, they will be pivot into many dummy variables. The number of dummy variable depends on the number values of your code under this categorical variables minus one. The one variable that was minus is the base line, all the left dummy variables of this categorical variable will take the base line variable as the reference.

Here is an explanation of dummy variables "marital_statusLiving common-law"(x4):

Under the variable "marital_status", we have 6 useful categories (exclude the NA category), and from x4 to x8 five variables shown in the from, the baseline is "marital_statusDivorced", so if the "marital_statusLiving common-law" is 1 means the respondent's marital status is Living common-law, if the "marital_statusLiving common-law" is 0 means the respondent's marital status is otherwise.

In this prediction model, b4 = 0.340159 means when respondent's marital status changed from Divorced to Living common-law, the respondent's feelings of life will increase 0.340159 if in the case of other conditions remain unchanged.

b4~b8 means when respondent's marital status changed from Divorced to their corresponding status, the respondent's feelings of life will increase or decrease by b value if in the case of other conditions remain unchanged.

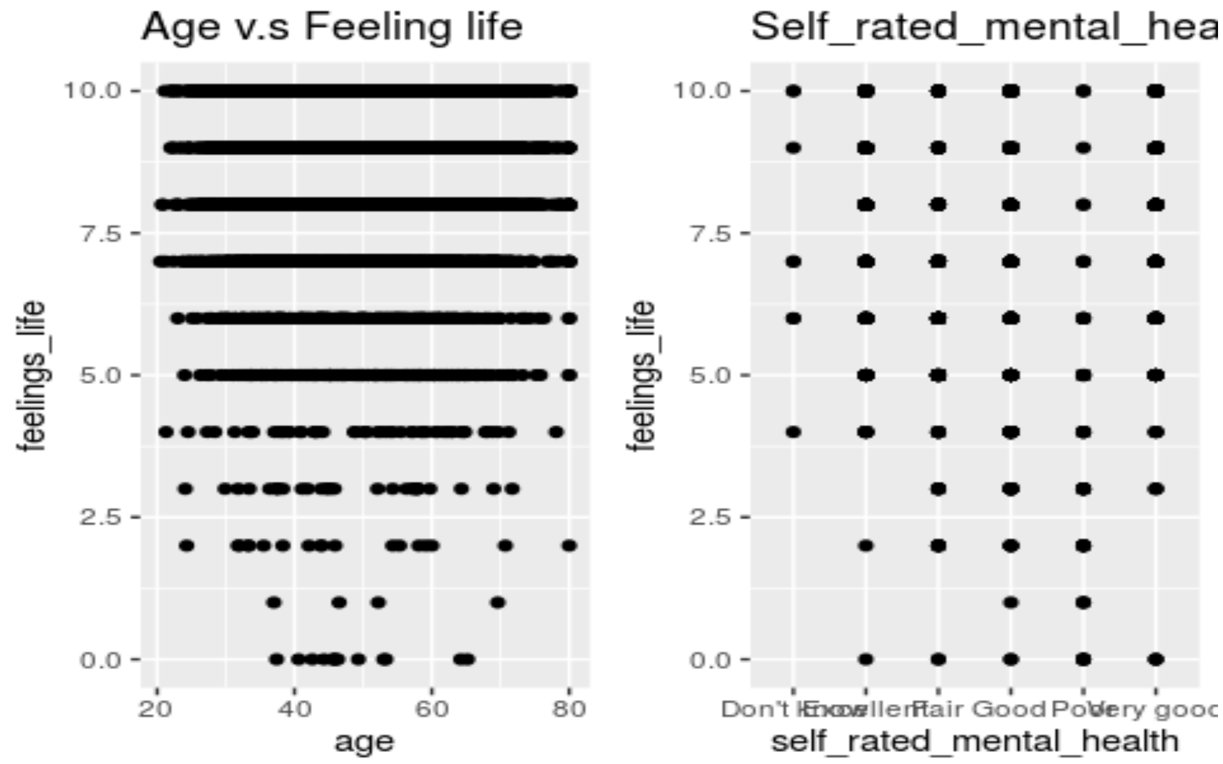
Another way to explain the coefficient value is : If we do b4 - b8 = 0.340159- (-0.266238) = 0.606397, it can be interpreted as when marital status changed from Living common-law to Widowed, the feelings of life will increase 0.606397 if other elements remain unchanged.

Similarly, b9~b14 means when respondents' education level changed from educationBachelor's degree to their corresponding education level, the respondent's feelings of life will increase or decrease by b value if in the case of other conditions remain unchanged. Same for the family income level, health level and mental health level. (b14~b29)

The first explanatory variable x1 is "age_at_first_birth" ,b1 means when the age of first birth increases every 1 unit, feelings life will decrease 0.015671 if other elements remain unchanged. The second explanatory variable x2 is "age", b2 means when the age increases every 1 unit, feelings life will increase 0.005035 if other elements remain unchanged.
The third explanatory variable x3 is "sexMale", this is the dummy variable of sex with the baseline "sexFemale". b3 means female's feeling of life is 0.110156 smaller then male if other elements remain unchanged.

However, we must evaluate the statistical significance of the estimates of our parameters, b, to see whether a particular x variable is making the useful contribution to the model. If the P-value of variable is smaller than 0.05, we will say this variable is significant to our prediction. From the form, we found the variable "age_at_first_birth", "age", "sex" and "marital_status" are significant to the prediction of feelings of life. And "self_rated_health" is the variable with the largest p-value in this model, the p-value of all the dummy variables are larger than 0.05 except the excellent level. So "self_rated_health" may not be considered as an element in prediction of feelings of life.

The R-square of this model is 0.2688, which means 26.88% of variation which is the feelings of life can be explained by this prediction model. Although the R-square is not large enough to show a good performance, this model can still be used to evaluate or explain people's feeling of life since our p-value is small enough to show the significant of this prediction model. And this prediction model is adequate for Canadians who are older than 15.



Results

Based on the model we have, we make scatterplots of feeling_life in respect to age, and feeling_life in respect to self_rated_mental_health.
[Fig1: age v.s. feeling life scatterplot]
On the scatterplot of feeling_life in respect to age, we observed that the data is relatively concentrated in larger feeling_life is satisfactory area, especially when age is also larger. For younger (near 20) and older (near 80), the satisfaction towards life is relatively higher. This makes us interested in whether there is a relation between age and feeling_life.
We construct a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, with x_1 being self_rated_health, and x_2 being self_rated_mental_health. This model yields a result, with R^2 = 0.2296, p-value < 2.2e-16. The regression model is as follows.
[model with mental and health variables only]   We use a nicer model instead. Eight variables are used to construct our multiple linear regression model, which are age_at_first_birth, age, sex, marital_status, education, income_family, self_rated_health, self_rated_mental_health. The first two (age_at_first_birth, age) are numerical variables, while the other six variables are dummy variables.

Our model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$, where y=feeling_life,
x_1 = age_at_first_birth, x_2 = age, x_3 = sex, x_4 = marital_status, x_5 = education, x_6 = income_family, x_7 = self_rated_health, x_8 = self_rated_mental health.
By the summary, $\hat{\beta}_0 = 6.978078$, this is our intercept, with standard error 0.615133, p-value <2e-16.
$\hat{\beta}_1 = -0.015671$, with standard error 0.002983, p-value = 1.53e-07.
This means with 1 unit increase in age_at_first_birth, holding other variables constant, feeling_life is expected to decrease 0.015671.
$\hat{\beta}_2 = 0.005035$, with standard error 0.001296, p-value =0.000104.
This means with 1 unit increase in age, feeling_life is expected to increase 0.005035.
The slopes for dummy variables sex, marital_status, education, income_family, self_rated_health, self_rated_mental_health are also listed in the summary above.
For dummy variable sex, slope of sexMale is estimated to be -0.110156. This means when sex is male instead of female, feeling_life is expected to decrease 0.110156, with p-value = 0.000454. When sex is female, this variable can just be neglected when calculating the estimated feeling_life.

For other dummy variable marital_status, the slopes and p-values are estimated by our multiple linear model as follows. From $x_4 to x_8$ we give the slopes for dummy variable “marital_status”, with baseline of “marital_statusDivorced”. From $x_9 to x_{14}$ we give the slopes for dummy variable “education”, with baseline of “educationBachelor’s degree”. From $x_{15} to x_{19}$, we give the slopes for dummy variable “income_family”, with baseline of “income_family$99,999 50 $125,000”. From $x_{20} to x_{24}$ are dummy variable “self_rated_health”. And from $x_{25} to x_{29}$ are dummy variable “self_rated_mental_health”
[b0,x1,x2-x29 estimate table]
The model gives residual standard error 1.262, with 7379 degrees of freedom. The R^2 = 0.2688, with p-value < 2.23e-16.

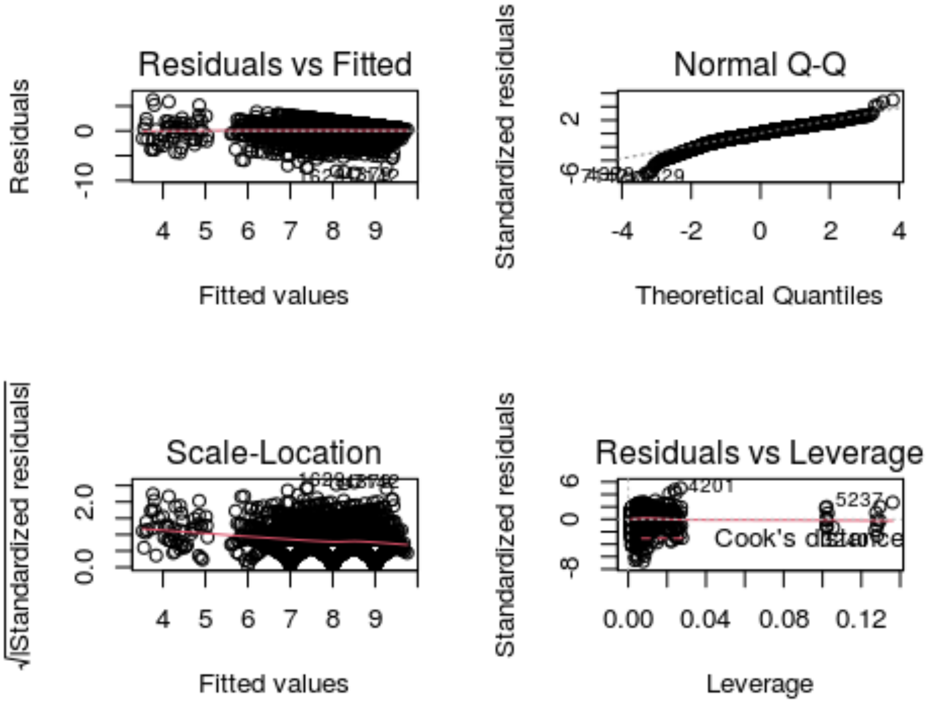
Discussion

Our model gives a R^2 = 0.2688, which indicates it is a statistically significant multiple linear model. With p-value < 2.2e-16, we reject the null hypothesis that there is no linear relationship on the model. Self_rated_mental_health has the largest estimate of slope, so it is the most influential on our y, feeling_life.
For numerical variables age and age_at_first_birth, the slopes has very nice p-values respectively. $\hat{\beta}_1 = -0.015671, with standard error 0.002983, p - value = 1.53e - 07.$ _{2} = 0.005035, with standard error 0.001296, p-value =0.000104. So the slopes of these

two variables are statistically significant in level of 95% significance.
For our dummy variables, when p-value is smaller than 0.05, we consider the respective slopes to be significant statistically. Those include sexMale (sex), marital_statusLiving common-law (marital_status), marital_statusMarried (marital_status), marital_statusSeparated (marital_status), marital_statusWidowed (marital_status), educationLess than high school diploma or its equivalent (education), educationTrade certificate or diploma (education), income_family$25,000 to $49,999 (income_family), income_familyLess than $25,000 (income_family), self_rated_healthExcellent (self_rated_health), self_rated_healthVery good (self_rated_health), self_rated_mental_healthExcellent (self_rated_mental_health), self_rated_mental_healthPoor (self_rated_mental_health). These variables are significant in linear regression, on the significance level 95%.
For some variables having p-values larger than 0.05, they are minor to our linear regression model. For example, marital_statusSingle in dummy variable marital status, never married has p-value to be 0.154716 > 0.05, which means we fail to reject null hypothesis that being single and never married is irrelevant to feeling_life levels. This probably because single persons vary in their satisfaction, and they are more based on their own values than basing on the relationships with others.
For dummy variable education, EducationCollege, CEGEP, or other non-university certificate has p-value = 0.165381, and EducationUniversity certificate or diploma below the bachelor's level has p-value = 0.276749, EducationUniversity certificate, diploma or degree above the bachelor's level has p-value = 0.584205. The three p-values are higher than 0.05, indicating there is no sufficient evidence to reject the null hypothesis for the three slopes, thus no evidence to suggest their linear relationship with feeling_life. This is probably because when education levels are high, people's satisfaction towards life are more related to something else other than their own educational level.
For dummy variable income_family, income_family$50,000 to $74,999 has p-value = 0.672565 > 0.05. The failure to reject null hypothesis suggest that middle-class people are less concerned on the effect of their income have on feeling_life.
For dummy variable self_rated_health and self_rated_mental_health, p-values are large for Good and poor and very good status. We figure this could be due to some different understanding of self_rated levels of health and mental health. The difference between definitions of good and poor is not very clear since they are both describing dissatisfaction somehow. Good and very good are also not divided clearly.

In conclusion, we present our model as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$, where y=feeling_life, x_1 = age_at_first_birth, x_2 = age, x_3 = sex, x_4 = marital_status, x_5 = education, x_6 = income_family, x_7 = self_rated_health, x_8 = self_rated_mental health. All the eight variables has linear relationship with feeling_life. What's more, in our model, self_rated_mental healthExcellent has the largest estimated slope, so it is the most influential variable on feeling_life.

Weaknesses



From the normal-QQ normal plot, we can see that the lower part deviates the normal line so that the standardized residual slightly violates the normal assumption.

Secondly, the plot of the residuals and fitted shows that most of residuals are around zero, but we can find some points lower right far from zero which violated assumption of the constant variance.

Next Steps

The bulk of the data we focused on is observational data. Yet it is not persuasive for observational data to derive any causation relationships. Although we take the model which is sufficient to indicate there is a linear relationship, it is hard to say there is a causation. The next step could be making experiments that has controlled variables to test the causation. However, this could involve some ethical problems.

References

1. 5.3 - The Multiple Linear Regression Model | STAT 462. (2020). Retrieved 18 October 2020, from https://online.stat.psu.edu/stat462/node/131/
2. gss.csv. (2020, October7) Author :Rohan Alexander and Sam Caetano
3. General Social Survey (GSS), Cycle 31, 2017: Family. Retrieved 18 October 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/doc/gss3x01.htm
4. Multiple Linear Regression. (2020). Retrieved 18 October 2020, from http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm