How does age or self rated life feelings affect a people's mental health

Group 49: Yingchen Tan, Xueqi Wang, XiWen Ran, Zezhou Han

2020, Oct, 19th

Abstract

In this context, the group of four people discussed the Canadian General Social Survey data list which collects Canada lived respondent's personal information during 2017. The group is defined as the relationship between respondent's age and self rated feelings rate that affect their mental health. Based on these data, we first used histogram, then used a logit regression model to make the discussion. The results show the effect in the variables which is the same as hypothesized: older people and people who rated higher on feelings life will get higher mental health scores.

Introduction

The project is made by YingChen Tan, XueQi Wang, XiWen Ran and ZeZhou Han. The project is based on the data of Canadian General Social Survey (GSS) and making the model to analyse our topic. Canadian General Social Survey is the data that collected people's personal and family information during 2017, which included family members, address location, income, etc.

The group is defined as the extent to describe one of the relationships in the dataset. We will discuss the topic of "How does age or self rated life feelings affect a person's mental health".

There is a variable called "self_rated_mental_health" which collects from each respondent, it represents the respondent's own self rated mental health score with positive side of "good, very good, excellent" and also negative side of "fair, poor". And all group members were interested in this variable. This variable is based on their self rated, which is different from other "real" data, you can directly see the value of height, the revenue or the family members, etc. We would like to know the reason how respondents rank their score. To discuss this variable, we choose variables "age" and "feelings_life". We will compute the correlation between "age" and "feelings_life" to confirm that it is independent or dependent on the ratio of "self_rated_mental_health".

To test the hypothesis —— the relationship between respondent's age and self rated feelings rate that affect their mental health. Respondents were randomly picked and answered with their personal information. They were asked to answer their age in number, self rated their life feelings rate from 1-10 and rate their mental health. Excluding N/A error variables, we also clean the data for variable "self_rated_mental_health", such that people rated "good, very good, excellent" means positive and "fair, poor" means negative. Based on these data, we first used histogram, then used a logit regression model to make the discussion.

We were hoping that we can find linear relationships between any variables and well reported them by observing the data, so making the logistic regression model, analyzing the result and discussion with conclusion will be our main works in the next steps. On this basis, we would consider analyzing the variables which could potentially affect people's mental health positively or negatively.

Data

By checking the dataset, we can say that the survey is from the Census of Canada. The target population for the 2017 GSS included all persons 15 years of age and older in Canada excluding the residents of Yukon, Northwest Territories, and Nunavut; and Full-time residents of institutions. In order to carry out sampling, each

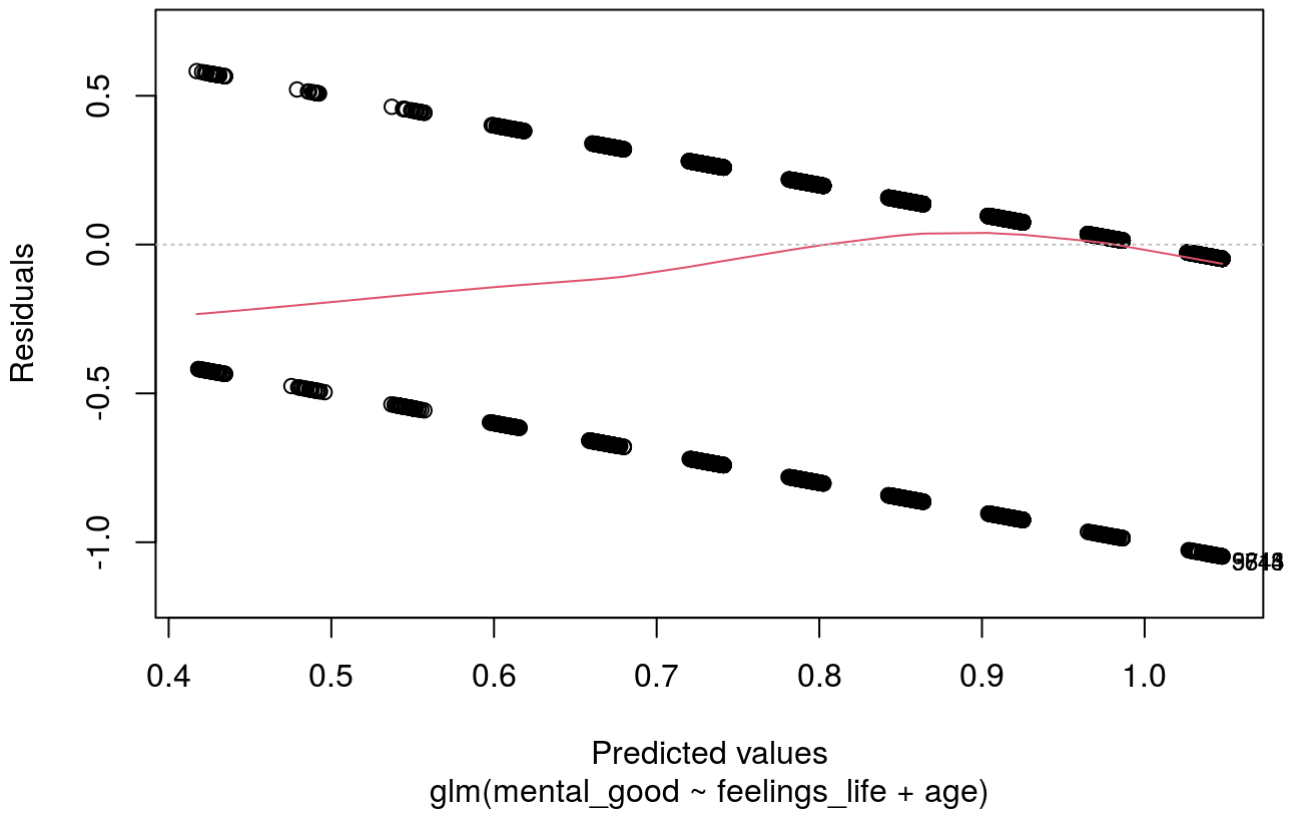
of the ten provinces were divided into strata(i.e geographic areas). Also, the non-CMA areas of each of the ten provinces were also grouped to form ten more strata, for a total of 27 strata.The payoff is this method takes the longest time cost compared to others. Since that's the Census of the country, I think there are no non-response situations during the experiment. But we still can make the hypothesis for if there are some people that make fake information or they just are not submitting their response Since we have some responses that are actually not available. For those not available responses, it might be that some independents didn't fill the survey fully or they don't know how to answer the question that we have provided. In this case, the weakness might be time costs and that also would cause bias with the final results. That would also cause Cognitive bias when the bias is large.

For this analysis, we have used the variable 'mental_health' as the dependent variable and for the independent variable, we choose the 'age' and 'feeling_life' to see the relationship between them. The x-values should refer to figure 2 and figure 3, that shows how the data are distributed. We consider some other variables to use such as age of children, number of children, etc. The strength of those two variables is that after the calculation of the correlation, the result tells us that those two are the best variables that would fit our analysis and that's why we only use those two variables as our predictors. Those two variables could tell us how Canadians feel about their life in 2017 and the distribution of their age. We collected those data by checking the survey and cleaning the whole data set, then we used the final version of the data for our analysis.

Model

Table1	estimate	pvalue	standard.error	t.value
Intercept	0.408200	<2e-16	9.963e-03	40.976
feelings_life	0.061310	<2e-16	1.064e-03	57.605
age	0.000329	0.000857	9.868e-05	3.334

Figure 1



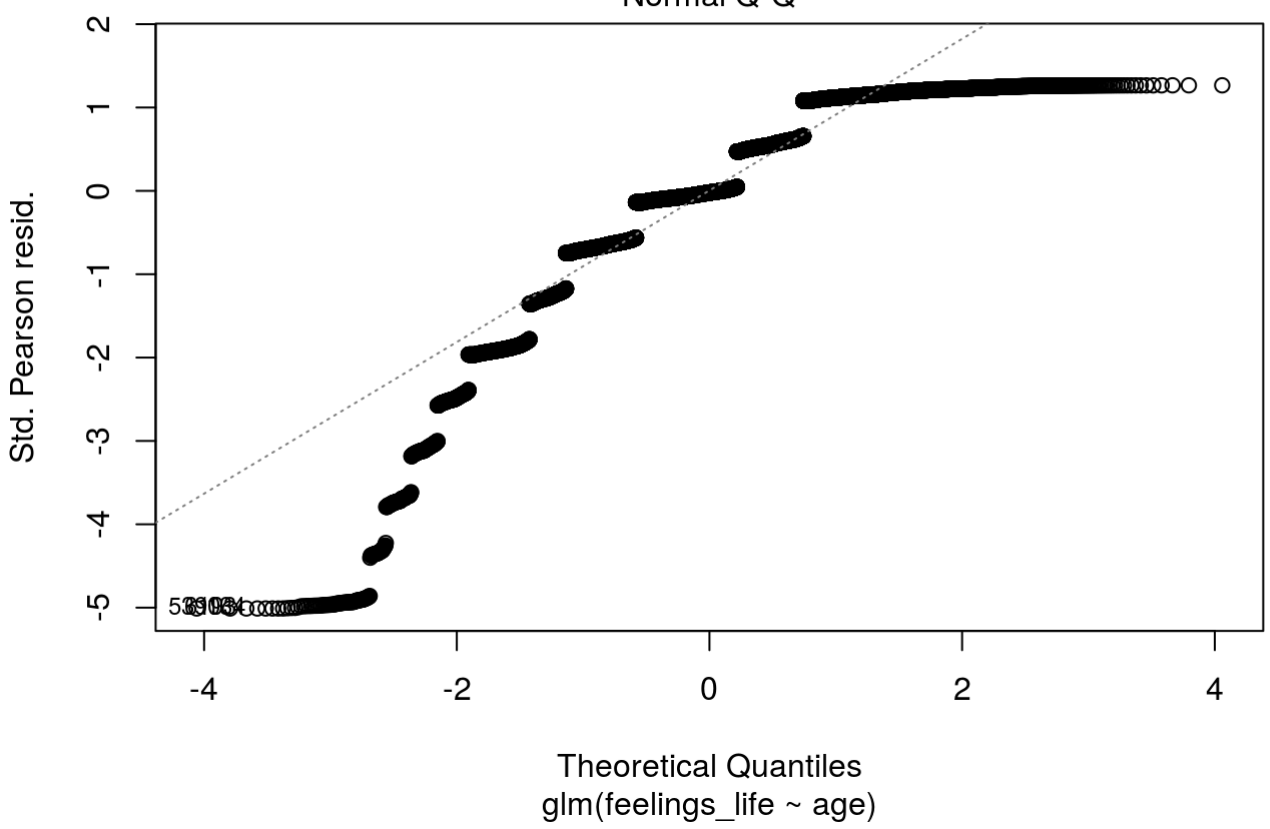
In this part,we use R to create a logistics regression model to find the probability of people's mental health in good status. We use feelings_life and age to be the explanatory variables. Since the age span of the sample is large, then age may be an important factor to affect people's mental health. Also, the variable feeling_life records their self rate to their life, which may directly show people's life satisfaction, so feeling_life should also be an important factor affecting people's mental health. Let p be the probability of people's mental health in good status, and feelings_life to be X_1 and age to be X_2, then the coefficient of X_1 is $\hat{\beta}_1$ and the coefficient of X_2 is $\hat{\beta}_2$. $\hat{\beta}_0$ is the intercept. From Figure 1 we can see the red line does not go over the data point range and it is not a straight line, which means using logistics regression assumption is appropriate.

From the summary table, we can see the $\hat{\beta}_1 = 6.131 * 10^{-2}$, $\hat{\beta}_2 = 3.29 * 10^{-4}$ and $\hat{\beta}_0 = 4.082 * 10^{-1}$, so the formula of the linear regression model is
$log(\hat{p}/(1-\hat{p}) = 6.131 * 10^{-2}X_1 + 3.29 * 10^{-4}X_2 + 4.082 * 10^{-1}$ for $X_1 \in [0, 10]$ and $X_2 \in [15, 80]$

Also, we can see the p-value for the null hypothesis that $\hat{\beta}_1 = 0$ and the p-value for the null hypothesis that $\hat{\beta}_2 = 0$ is 0.000857. Both numbers are very small numbers, so we can reject these null hypotheses. It means that age and feelings_life are relating to the probability of people's mental health in good status.

Table2	estimate	pvalue	standard.error	t.value
Intercept	7.8516643	<2e-16	0.0357851	219.412
age	0.0046723	6.97e-13	0.0006503	7.185

Figure 2



For the alternative model, we can create a linear regression model for age and feelings_life. The strength of this model is we can directly know how age and people's self rate to their life. Since the p-value for the null hypothesis that estimated slope equals to 0 is $6.97 * 10^{-13}$, which is much smaller than 0.05, then age and

feelings_life may have linear relations. However, from Figure 2, we can see the weakness is the QQ plot does not follow normal distribution, which means using linear regression model may be not suitable for age and feelings_life.

Results

Figure 3 - Histogram of life feelings

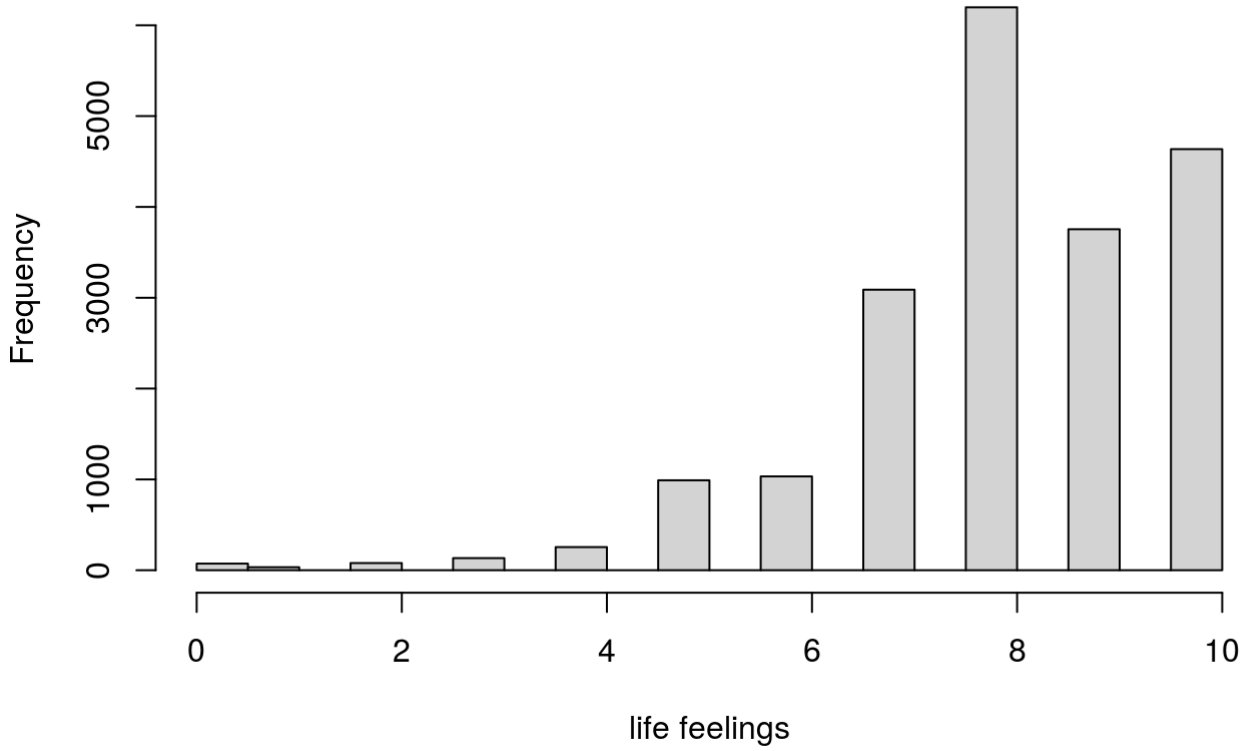
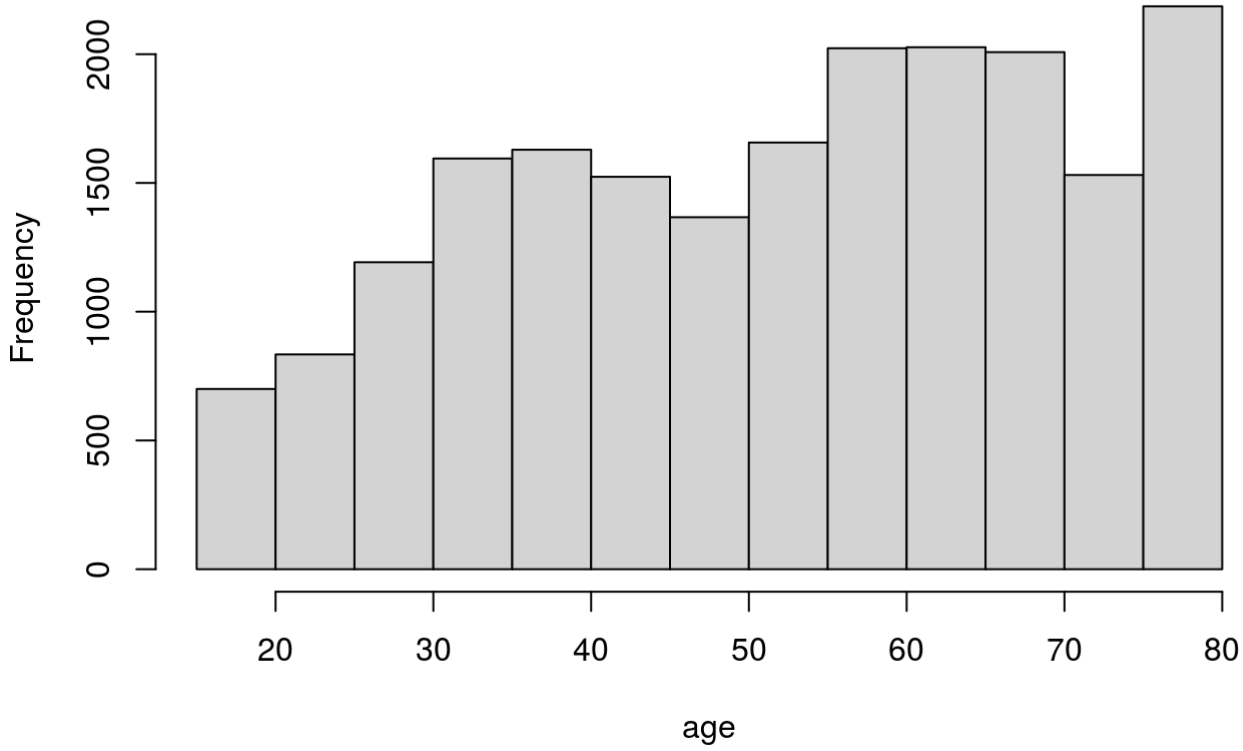


Figure 4 - Histogram of age



The Figure 1 of Residuals vs. Fitted plot displays the linear relationship between variables, which demonstrates equally spread residuals around a horizontal line without distinct patterns. The data points were simulated in a way that met the model assumptions precisely, indicating that the model was quite accurate. From Table 1 from our logistics regression model, we can see that the variables "feelings_life" and "age" are both statistically significant because the corresponding p-values are less than the significance level of 0.05. In short, "feelings_life" and "age" are two variables that relate to the probability of individuals' mental health in good status. When the scale of individuals' feelings about life becomes larger, individuals' mental health status would more likely be good. On the other hand, age is also a significant contributor to the fairly good status of mental health.

The Figure 3 of histogram of the age of respondents shows that most of the respondents belong to the age brackets of fifteen to eighty years. The Figure 4 histogram of individuals' feelings about life as a whole explores a scale of 0 to 10, where 0 means "Very dissatisfied" and 10 means "Very satisfied." It can be seen that most of the data values in the histogram are in the range of 5 to 10. This leads to the conclusion that a large number of individuals are satisfied with their life.

After looking at all figures and the table of our model, we can conclude that "feelings_life" is the variable that has more influence on the status of individuals' mental health, compared to the "age" variable.

Discussion

The result of our statistics, figures and the table, targeting the population for the 2017 GSS including all individuals fifteen years of age and older in Canada, shows that maintaining a good status of mental health can be affected by individuals' feelings about life as a whole. We find similar trends in age, and we can access the relationship between age and mental health status, which older people are more likely to achieve and sustain a state of good mental health. To avoid caveats, we clean up the inaccurate data, which is not appropriate for the purpose of getting the correct model. We keep the data values which have a strong influence on our observation results.

From the book "Statistical Rethinking: A Bayesian Course with Examples in R and Stan" by Richard McElreath (2016), the small world and the large world always provide the opportunity for a conversation between model and reality. The small world represents the model itself, while the large world is what we hope to deploy the model in, which is behind the presentation of the model's data. There exist challenges in our statistical modeling. Although we attempt to minimize any uncertainty related to our data, there are still weaknesses that we must improve in the future.

-Weaknesses

One of the weaknesses of our work is that the model is limited by the quality of the data. Besides, the data for those individuals who answer "Don't know" to particular questions cannot contribute to the analysis. As a result, when the data is not fulfilled, the result will be biased as well. For future improvement, we would consider analyzing more variables that could potentially affect people's mental health positively or negatively. We should also make sure that the most significant variable in determining the mental health status of individuals would be provided. Thus, it is possible for us to analyze the difference between various influence factors.

-Next Steps

Since the data collected people's personal and family information during 2017, there should be some different questions to explore in 2020. Based on our current results, it is essential to conduct a follow-up survey, which helps us to estimate the accuracy and precision of the results obtained. We would include the questions, such as asking individuals' feelings about life or their feelings in general. In addition, there would be questions that relate to the effect of COVID-19. The important variable we are going to measure is to what extent, COVID-19 influences individuals' day to day life, given a scale of 0 to 10.

References

2017 General Social Survey: Families Cycle 31, PUMF-Data Dictionary, February 2020. Retrieved October 19, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/index.htm (https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/index.htm)

Government of Canada, Statistics Canada. General Social Survey on Family (Cycle 31). Retrieved October 19, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm (https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm)

McElreath, R. (2016). Statistical rethinking a Bayesian course with examples in R and Stan. Boca Raton: CRC Press, Taylor & Francis Group.

RStudio [Computer software]. Retrieved October 19, 2020, from https://rstudio.cloud/projects (https://rstudio.cloud/projects)

Appendix

This file was made using the R markdown package: tidyverse, dplyr, tinytex. All code used in this paper can be accessed from within the code blocks of the markdown document. All the methodology we have developed are learned in STA304H1F of University of Toronto (St. George campus).