Impact of age and gender on life satisfaction

Xinran Gu

2020/10/19

Abstract

This report investigates how age and gender affects life satisfaction. A multiple linear regression model and other graphs are used in the report. The multiple linear regression model could be represented as$\hat{y} = 7.8803 + 0.0046x_1 - 0.0595x_2$. This indicates that age and gender does have influence on life satisfaction.

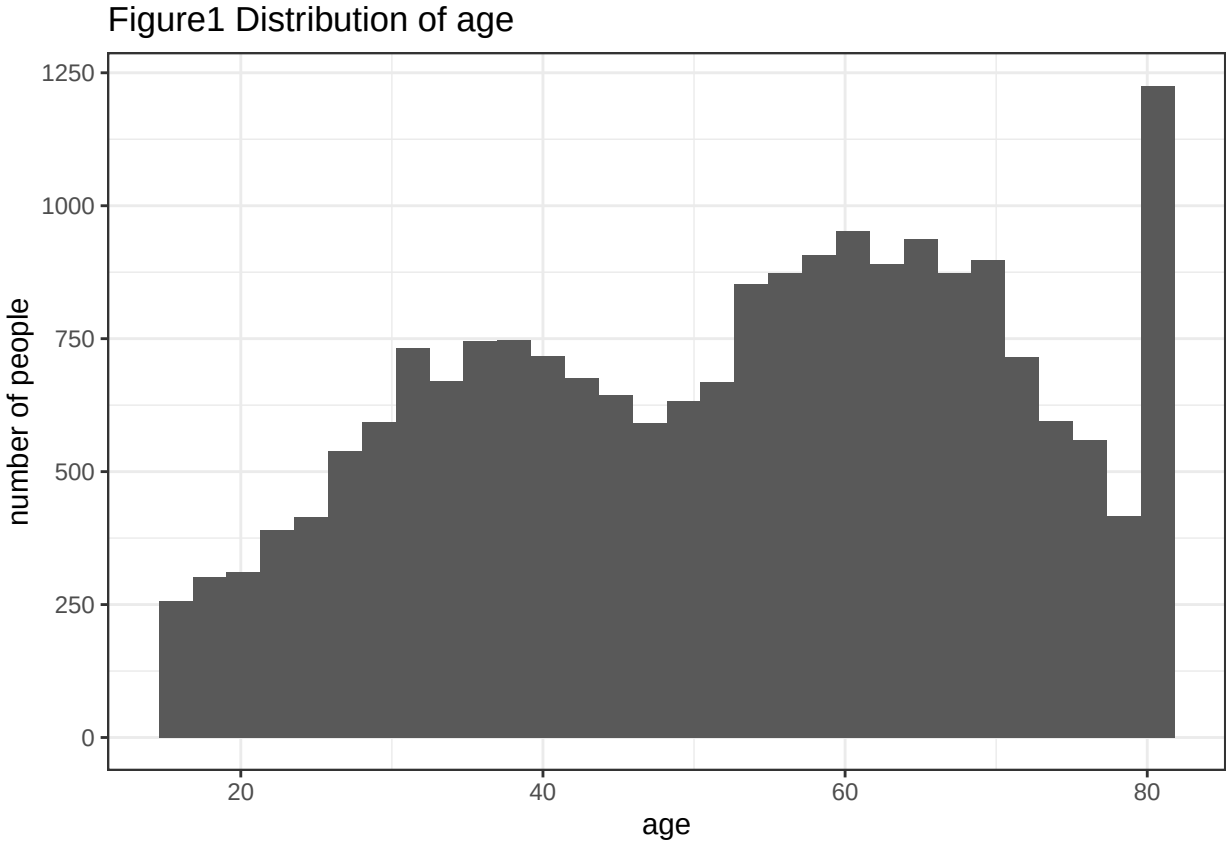
Introduction

Life satisfaction is important as it not only affects a person's happiness but also the long-term well beings. The response variable of the report is feelings about life. The two explanatory variables of the report are age and gender. The reason why I would like to use age and gender as explanatory variable are listed below. First, different age means a person is in different phase of life. For instance, age 15 indicates he/she is in school and age 30 indicates he/she enters society. I believe these differences will have significant influence on one's life satisfaction. Second, different gender also indicates a person will experience different situation such as pregnancy. Before I analyze the data, my hypothesis was that the satisfaction level of life will decrease as age increases and female will be more dissatisfied. However, the results of analysis is totally opposite. The model shows that the satisfaction level of life will increase as age increases and female is actually more satisfied. The report will first discuss the data that I used. Then I I will use multiple regression model and multiple graphs to investigates how age and gender will influence life satisfaction. After that, I will talk about the weakness of the analysis and what can be improved in the future. Code and data supporting this analysis is available at: https://github.com/LiviaSta/Assignment2

Data

The data that I used in this report is the 2017 General Social Survey(GSS) on the family. According to the user guide of 2017 General Social Survey, data collected by General social Survey comprised two components.The core content is designed to measure changes in society related to well beings. The classification variables such as age and gender which I used in this report will help delineate population groups. The target population of GSS includes all persons who are 15 or older in Canada.Yet, it excludes full-time residents of institution and residents in Yukon,Northwest Territories and Nunavut.The frame population consists of lists of telephone number in use available to Statistics Canada and lists of address registered within the ten provinces. Each of the ten provinces were divided into strata and then a simple random sample without replacement was performed.Respondents who are 15 years old or older are eligible and were interviewed through telephone. If the phone call was not reposed, interviewers were made numerous call backs. The strength of the data is that it used stratified random sampling, so it has a acceptable sampling variability. Yet, one significant drawbacks of GSS is that it is hard to generalize the results to the whole country. The sample size of GSS is 20,602 which is relatively large compared to other survey. However, the estimate population of Canada is 38,005,238. Moreover, there exists some biases in the data. For example, age and gender are the variables that I used in this report. By looking at Figure 1 below and Figure 4 in the appendix, I

notice that the percentages of female and older people are higher. This will definitely have impact to the results of the analysis and decrease validity.



Model

The model that I used in this report is multiple linear regression model and the model was created by the lm function in r. This model allow us to investigate the linear relationship between our response variable and two different explanatory variables.Even though gender is a binary response variable, I feel it is not enough to predict the life satisfaction level based solely on gender. Hence, I did not choose Logistics regression model. The reason why I choose age instead of age group is because I think a continuous variable could show how the increase of age affects life satisfaction better. The second explanatory variable of the report is gender which is a categorical variable. This is because the values in gender are male and female. The lm function in r will automatically compute the categorical variable, and I will discuss this further below. The linear regression model between feeling about life and age and gender is represented as : $\hat{y} = 7.8803 + 0.0046x_1 - 0.0595x_2$. In the model $\hat{y}$ is feeling about life, x_1 is age and x_2 is gender. We also need to notice that in this model we use "sexMale" as a coefficient, it means we use 1 to represent male and 0 to represent female. Hence, the model representation for male is $\hat{y} = 7.8208 + 0.0046x_1$ and the model representation for female is $\hat{y} = 7.8803 + 0.0046x$.The specific analysis of the model will be in the results and discussion sections.

Results

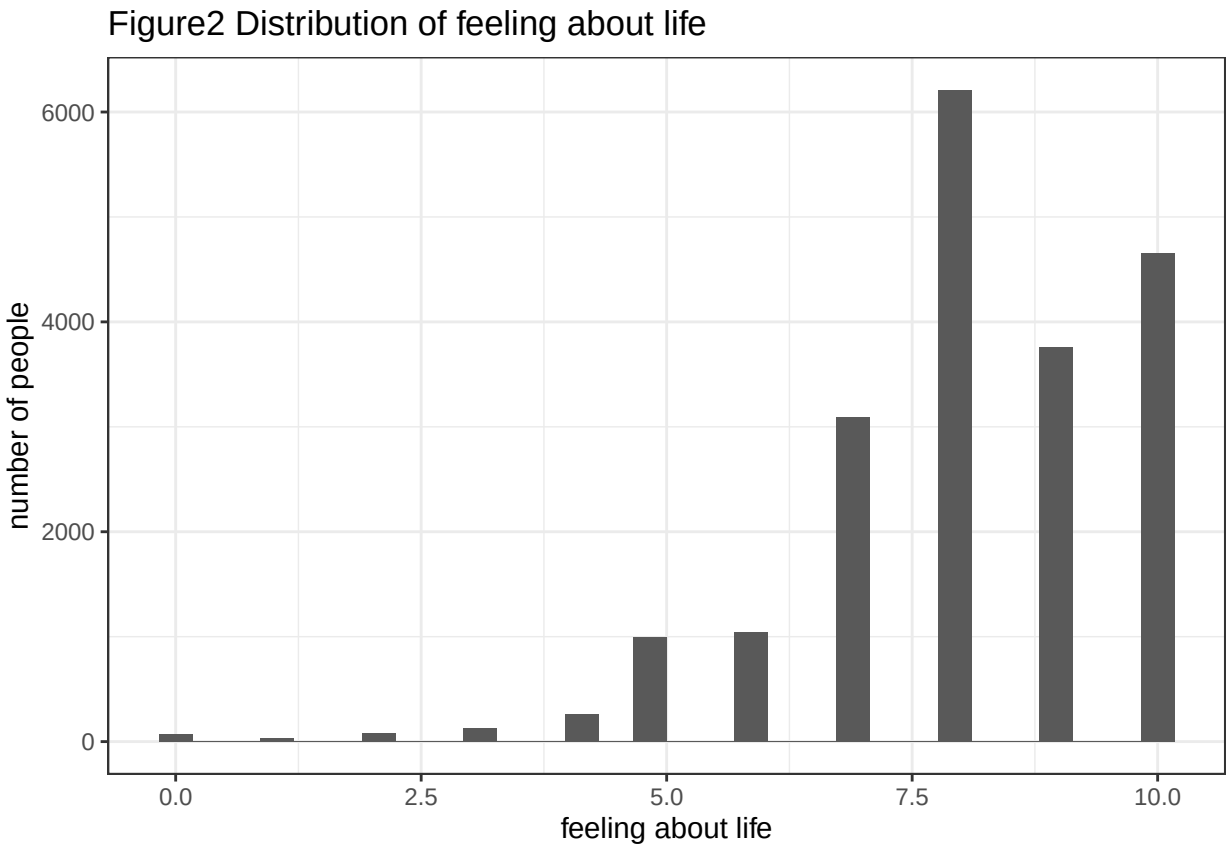


Figure 2 is a histogram of life satisfaction level.This histogram has a life-skewed distributions, it has a tail extends to the left while most values cluster on the left.

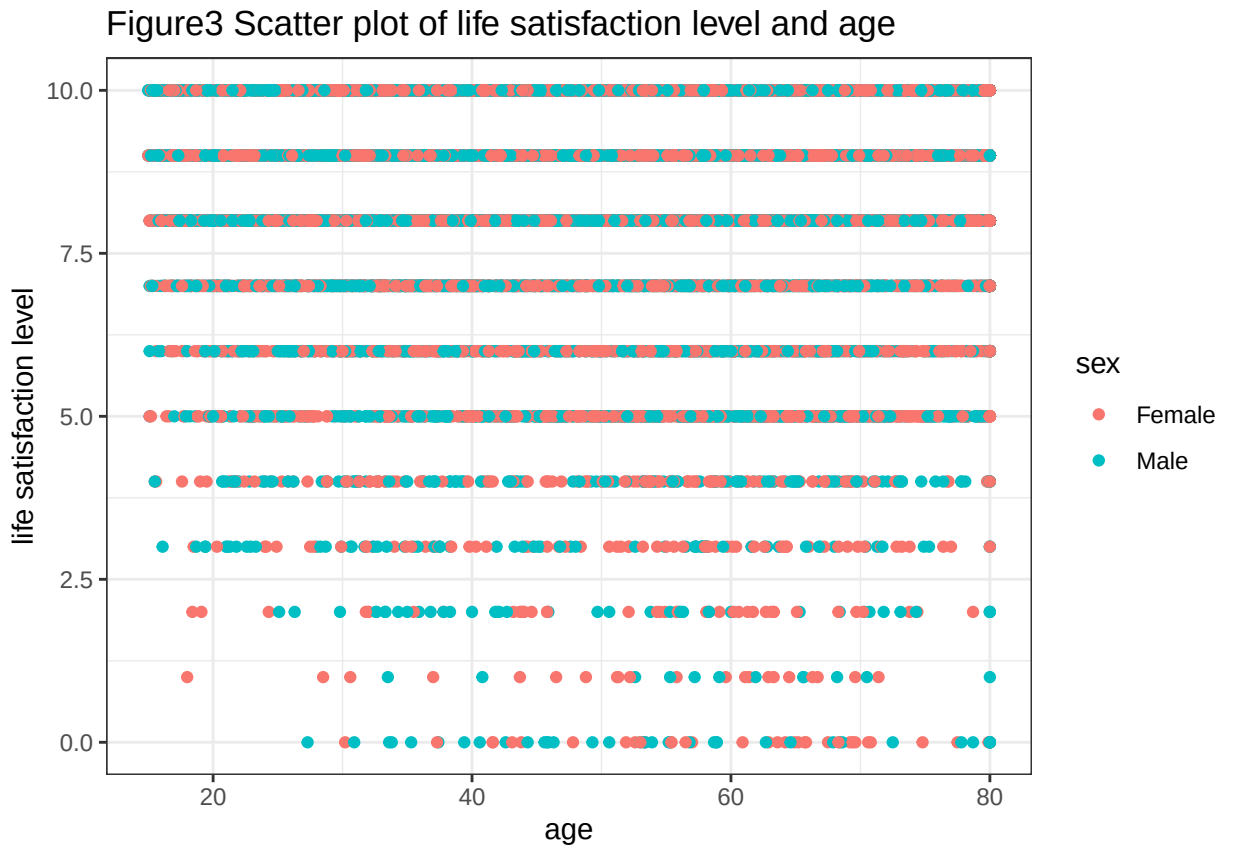


Figure 3 is a scatter plot of life satisfaction level and age. By looking at the graph,it seems like older people tend to have lower life satisfaction. However, if we think about Figure 1, it shows that there are more older people in the sample as the first quartile of age equals to 37.3. The median and mean of sample age are 54.1 and 52.11 respectively. .

In order to find the linear relationship, it is necessary for us to analyze the results of the results of the multiple linear regression model.The intercept estimate $\hat{\beta}_0$ in the linear regression model is 7.8803, it is the average value of life satisfaction level when age and gender are equal to zero.The slope estimate $\hat{\beta}_1$ in the linear regression model is 0.0046. This suggests when age increased by 1, the life satisfaction level will increase by 0.0046 for both female and male on average. $\hat{\beta}_2$ in the linear regression model is 0.0595 which is the difference in average of life satisfaction between male and female.

Discussion

By interpreting Figure 2, we know most of people rated their life satisfaction level around 8. This also proved by The mean and median of the data of feeling about life, which are 8.094 and 8 respectively. Also the first quartile of life satisfaction is 7 which is high value. We need to keep this mind that only 25% of people rated their life below 7. By interpreting Figure 3,the variance is 314.725 which is really large. This large variance indicates that the data are very spread out from the mean which could be seen from the high number of people who is around 80. To further investigate the multiple regression model, we need to take a look at the standard errors. $se(\hat{\beta}_0)$,$se(\hat{\beta}_1)$ and $se(\hat{\beta}_2)$ equal to 0.0377, 0.0007 and 0.0231 respectively. It is good for us to have small standard errors when we compare them to the estimators. The small standard errors suggest that our estimated values are close to the actual values. After that I would like to do hypothesis tests at significant level of 5%. First, for the slop,The null hypothesis is $\beta_1 = 0$ and the alternate hypothesis is $\beta_1 \neq 0$. Since the p value 1.14e-12 is definitely smaller than 0.05, we reject the null hypothesis. This means the relationship between life satisfaction level and age is significant.Then, for whether there is a average

difference between male and female,we write the null hypothesis as $\beta_2 = 0$. The alternate hypothesis is $\beta_2 \neq 0$.As the p value equals to 0.0102 which is smaller than 0.05, we reject the null hypothesis. Hence, we said the average difference between male and female on feeling about life is significant.

Weaknesses

One significant weaknesses of the survey is that the sample size is hard to generalize to the population which is the whole country.Moreover, even though the stratified sampling allows geographical variability, there are still exits biases on gender and age.Also, in the survey, the question was asking to use a scale of 0 to 10 to rate their life as a whole right now. The first problem of this question is that feeling is really subjective and hard to express quantitatively. One person's very dissatisfaction might equal to otter's.Second, the question asked the participants' feeling at that moment. It is possible that there exists participant error as their mood might affect their answers at that time. The weakness of the analysis will be there are many other factors that might have influences on life satisfaction other than age and gender. Therefore, the results of the analysis probably could not fully explain why life satisfaction changes.

Next Steps

If we do a follow-up survey, we could also use stratified random sampling. Yet, this time we should use age group and gender as strata. This will increase the validity of the results. In order to improve the model, we need to add more variables to it. For example, the number of children and income could also affect life satisfaction.

References

1.Alexander, R. (2020, May 17). Telling Stories With Data. Retrieved October 16, 2020, from https://www.tellingstorieswithdata.com/ 2.Government of Canada, S. (2020, October 16). Population estimates. Retrieved October 16, 2020, from https://www150.statcan.gc.ca/n1/en/subjects/population_and_demography/population_estimates 3.General Social Survey Cycle 31:Families Public Use Microdata File Documentation and User's Guide. (2020). Ottawa: Statistics Canada. 4.2017 General Social Survey: Families Cycle 31 Public Use Microdata File. (2020). Ottawa: Statistics Canada. 5.Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org 6.Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686. ## Appendix 1. bar plot of gender Figure 4 is a bar graph that shows the number of female and male. It is important for us to notice that there are more female participated in the survey, so there might exist some bias in the results of the data.

Figure4 Gender difference