Life Satisfaction Analysis of Young Adults Based on Differences of Province, Age and Personal Income

Yuehao Huang, Suran Wu, Siyu Chen, Lingyi Li

Oct.16, 2020

Code and data can be accessed via github at https://github.com/EroSkulled/STA304/ , licensed under MIT.

Abstract

We use the multiple linear regression technique to analyze the impact of personal income, age and regional difference on self-rated feeling of life. We make the hypothesis that the younger and higher income the respondents have, the higher they will respond on the feeling of life number scale and people from major cities will generally have a more satisfied life. However, our model gives a prediction that people with higher ages will slightly have a lower satisfaction of life and only those with the lowest annual income(less than $25k) would have a negative impact on their feelings of life. The impact from regional differences is only seen in three provinces.

Introduction

This report investigates the relationship of individuals who are under 35 feeling of life and key factors that may affect how one rates his life, namely, age, province of residence and personal income. Will moving to Quebec from Ontario improve your self-feeling? Will a high income contribute to your happiness? Do people prefer young age (therefore, no burden) or do people tend to be happier with wisdom and experience from adultness.

A multiple regression model would be used to analyze their relationship since there are multiple explanatory variables to predict the response variable. Plots will be made to visualize and show that the multiple regression model is appropriate in the Data section. Data is collected and obtained by Statistics Canada, please refer to the Data section for more comments and notes about data.

In the Result and Discussion sections,results from the model will be presented and interpreted. In addition, a conclusion will be drawn based on analysis from the data and model.

The last part of this analysis consists of weakness of the model and data, and next steps, suggests what subsequent and potential work can be done for improvement.

Data

The data we used are self-rated feelings of life, province, age and income. For age, we are interested in young adults that are below 35 and they are obtained from: 2017 General Social Survey: Families Cycle 31 from Statistic Canada, see reference. Method used to carry out the survey consists of 2 steps, first, a stratification of 27 strata was performed based on geographic area1. Then, a simple random sample without replacement of records was performed within each strata. So a stratified random sampling was used. Stratified random

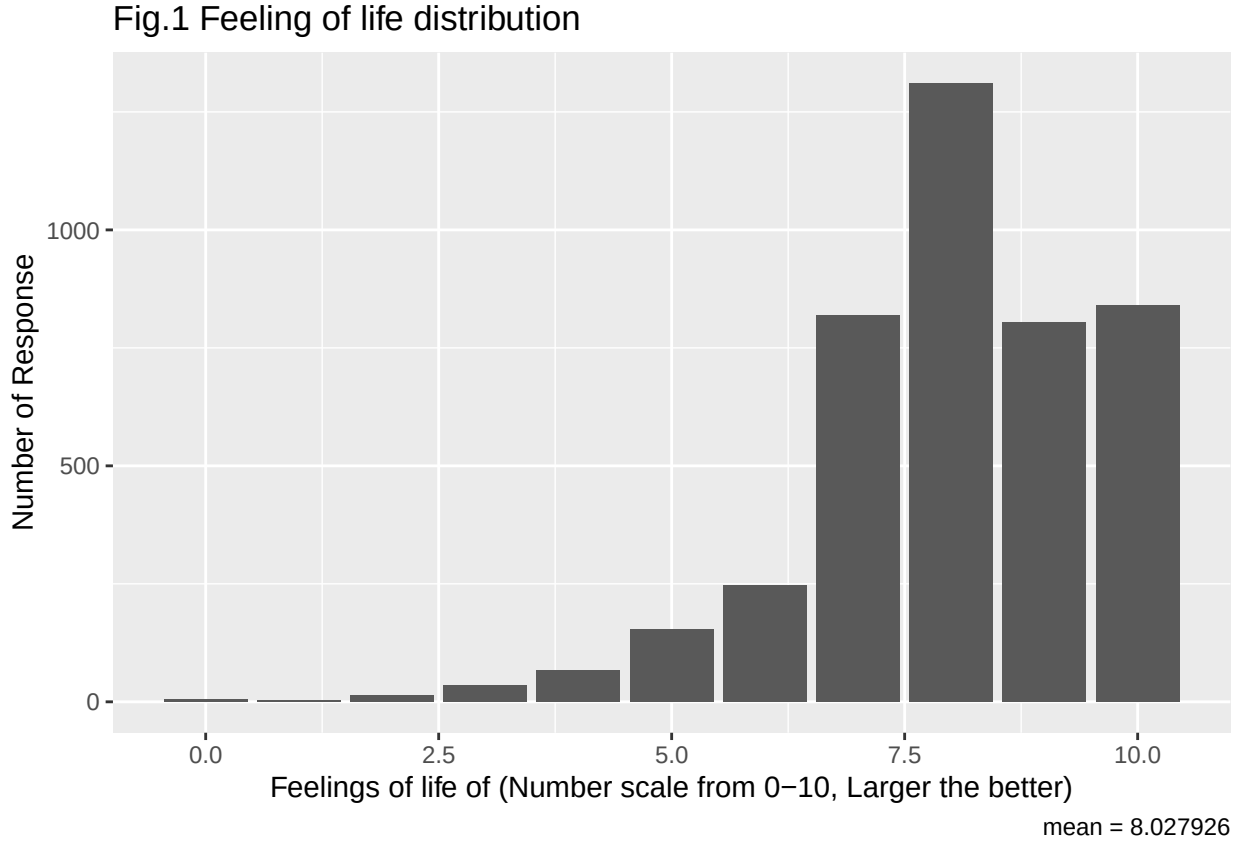
sampling is a probability sampling technique that slices population into homogeneous groups, then random sampling of each strata is used.

Advantage of performing a stratified random sampling is quite clear. Stratified steps ensure minimizing sample selection bias and decreasing difficulty accessing full lists of population. Those are 2 disadvantages of simple random sampling on its own. And simple random sampling within each strata still provides a simple approach that lacks bias within each strata. Some cons are notable as well. For example, to perform such sampling, retrieving frames and deciding how stratified steps work can be very capital and time consuming. How to slice population strata is another difficulty that requires professional analysis. Statistic Canada needs to deploy quantitative specialists to finish those steps.

For self-rated feelings, it consists of 11 scales from 0-10 besides standard non-response options. Strengths like providing direct feedback of subjective satisfaction exist. Some drawbacks are clear. It lacks scale to determine each interval. Additionally, It asks 'how do you feel about your life as a whole right now?' Some may answer a false high score because they just had a happy event and would decrease their rate after a couple of days. Other than that, the feeling of life as a whole is too comprehensive and vague and can be difficult to answer.

For province and age, besides standard non-response options, it consists of options 10 provinces for province of residence and age is filled with numeric answers rounded to one decimal place. And again, we are interested in people who are below 35 years old. Since age and province are relatively objective answers and hence no clear strengths and drawbacks will be discussed.

The distribution of self-rated feelings with age under 35 are as the following plot:



As we can see most respondents are within a range of 7-10 and a mean of 8.027926, meaning most people are quite satisfied with their life.

For income, besides standard non-response options, it is categorical with an interval of 25000 annually. It consists of 6 answers of below 25000 all the way to over 125000. The strength is that it uses tax to conclude

personal income rather than let respondents choose on their own. It effectively prevents people from giving false information. Potential drawback would be tax evasion, and therefore the concluded income from tax may underestimate the actual personal income.

Method used to carry out the survey consists of 2 steps, first, a stratification of 27 strata was performed based on geographic area1. Then, a simple random sample without replacement of records was performed within each strata. So a stratified random sampling was used.

The population of this survey includes all people that are older than 15 years old, but excludes residents of the Yukon, Northwest Territories and Nunavut and excludes full-time residents of institutions. The survey frame consists of 2 components. First is lists of telephone numbers (both landline and cellular) available to Statistics Canada, then, the Address Register which is all dwellings within the 10 provinces. Sample were respondents who were randomly selected from each eligible household and accepted the survey. Target sample size is 20000 and 20602 samples were obtained.

Model

We use the package lm function embedded in R Studio to create this multi-linear regression model. The model is defined as follow:

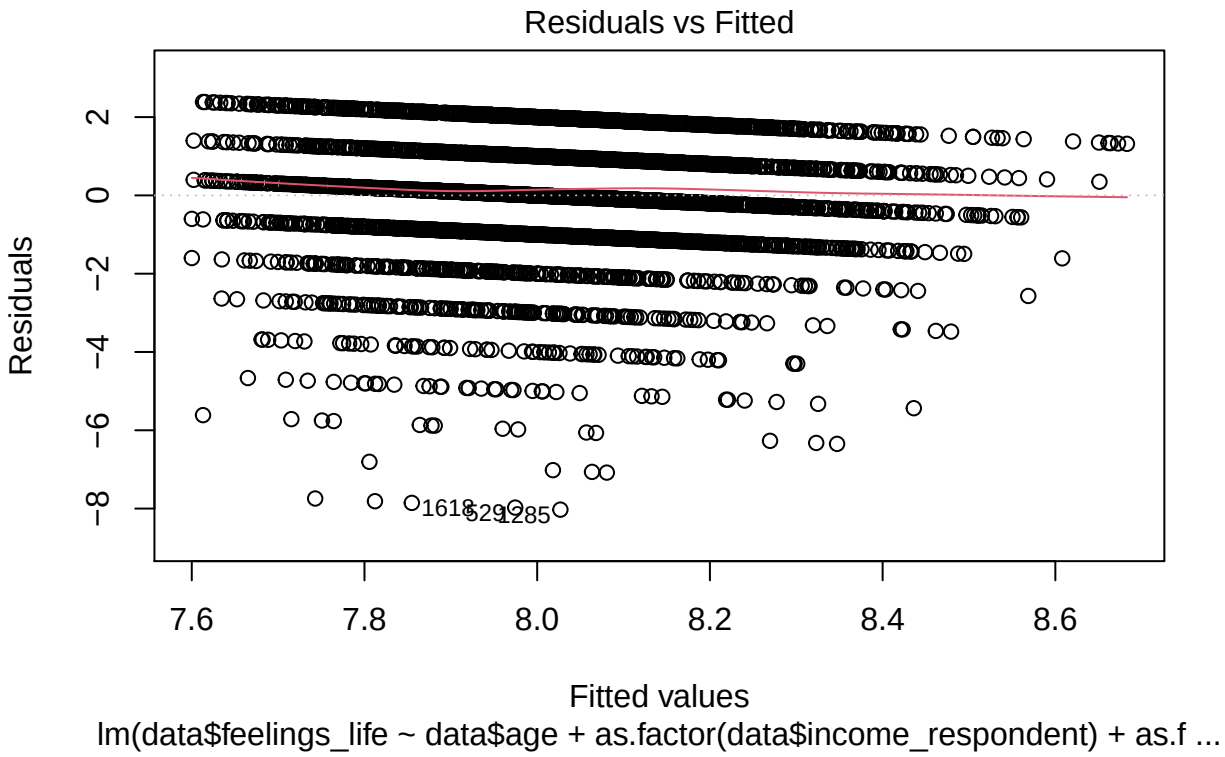
$$Feeling.of.life = 8.75055 - 0.02156 \times X_{age} - 0.37600 \times X_{Inc<25k}$$
$$+0.25768 \times X_{NB} + 0.28392 \times X_{NL} + 0.18857 \times X_{QC} + \epsilon_i$$

Note the notation is explained as the following: X_{age} represents the Age of the respondent, and $X_{Inc<25k}$ represents people who falls in the category that has income less than 25,000 dollar. The rest variables, such as X_{QC}, stands for if the respondent lives in the province of Quebec. The subscripts indicate the abbreviation of the province names respectively. For province and income, if respondent falls into the correct category, X = 1. Otherwise it equals to 0 as it is functioned as an factor(categorical variable). For example, Say an adult with Age 28 lives in Quebec has an annual income of 47,500 dollar. Then the predicted feeling towards life is:

$$8.75055 - 0.02156 \times 28 + 0.18857 \times 1 = 8.33544$$

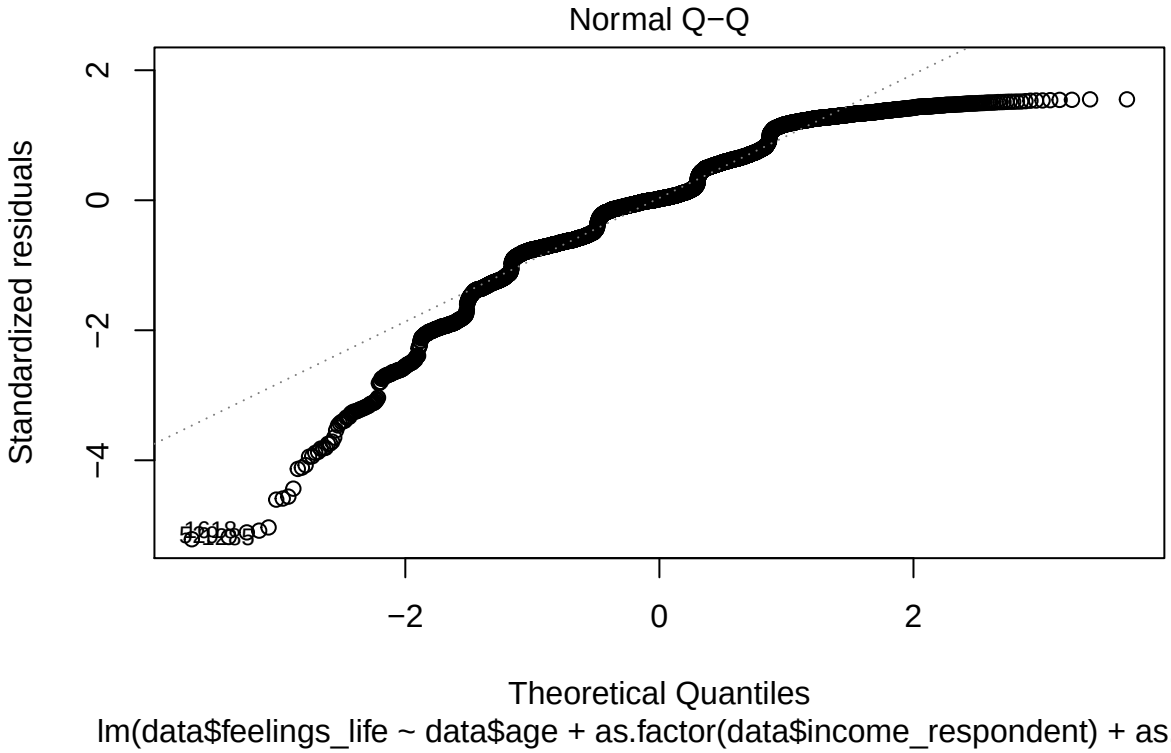
The limitation of this model is that we do not have enough sample within some of the province and hence can result in extreme expectation. We also would anticipate it to be not normal distribution as the model consists of many categorical variables.

Fig.2



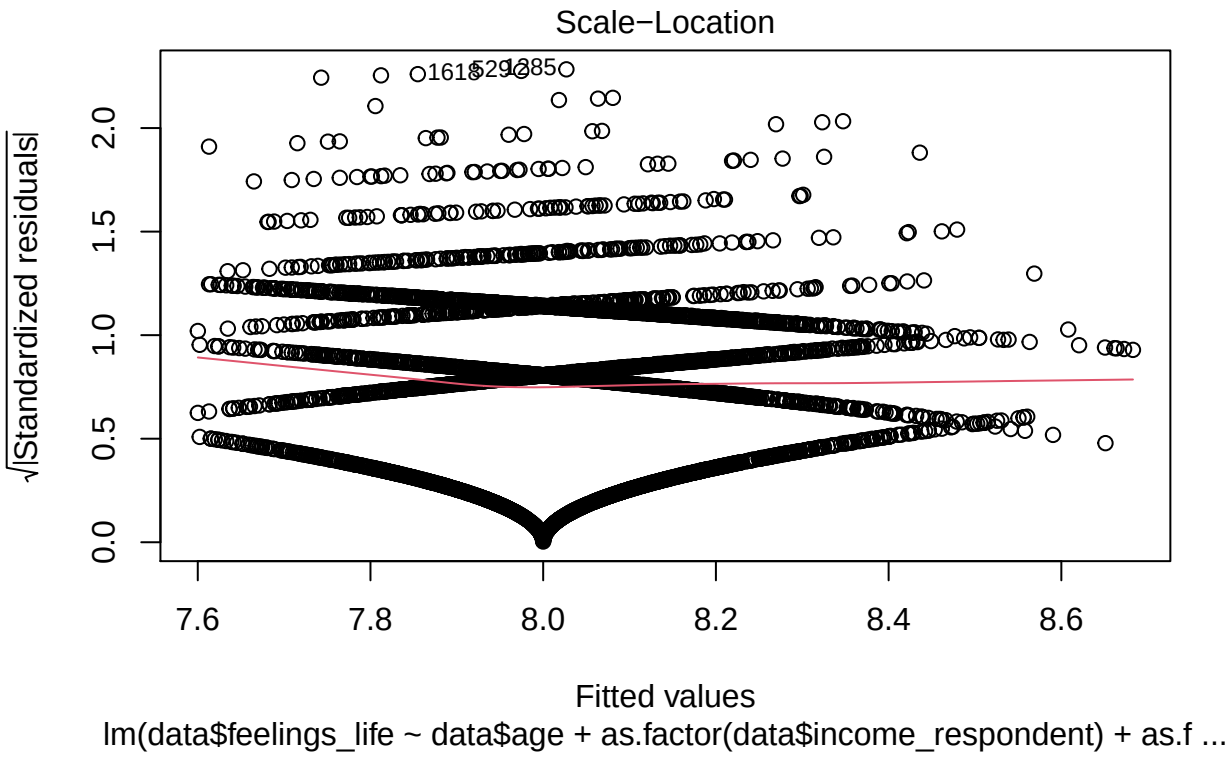
In Figure.2, the Residuals vs Fitted plot shows if residuals have non-linear patterns. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships. However, in our case, there is clearly a pattern along the red line, therefore the residuals do not have a non-linear pattern.

Fig.3



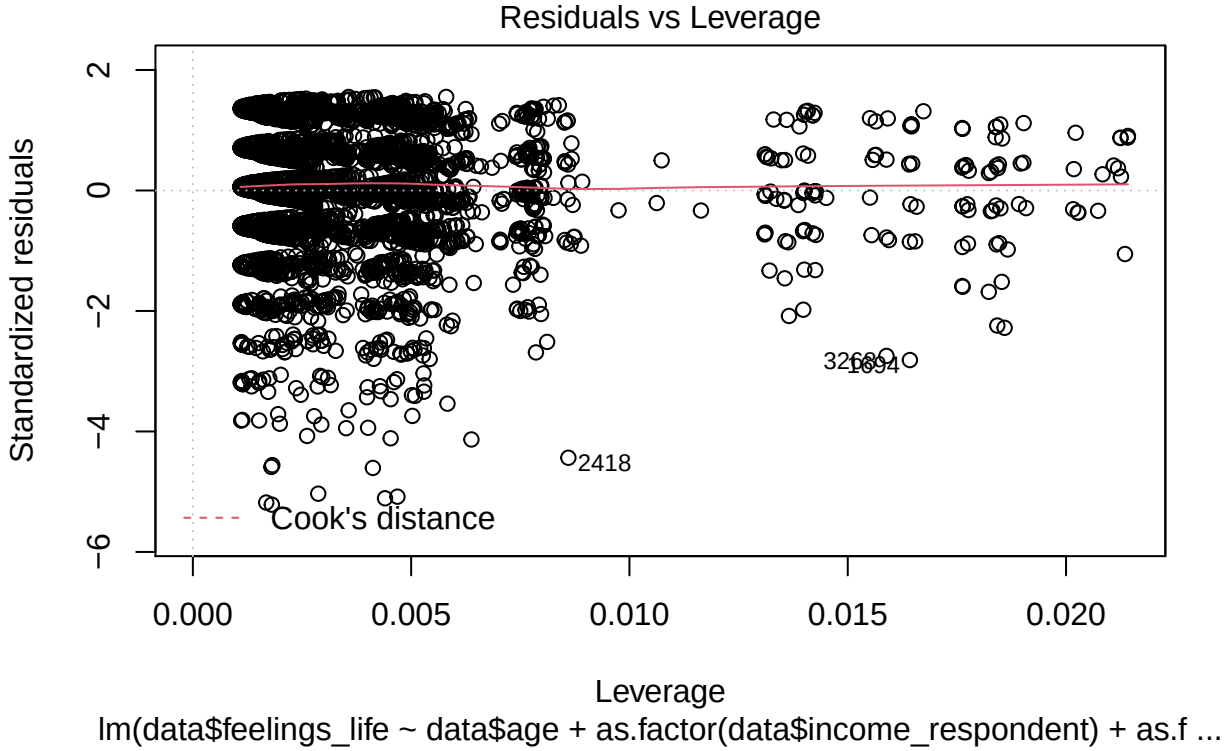
The normal QQ(Figure.3) plot shows if residuals are normally distributed. If the residuals follow a straight line well then it is good. In the normal QQ plot I have, I will say it is not a good model. Thogh still valid, the data is not normally distributed and meaning we could not get a precise result using thie model for prediction.

Fig.4



In the Scale-Location plot(Figure.4), there is a 'horizontal line' and the points are not randomly distributed along the line. So the assumption of equal variance is not achieved.

Fig.5



The Residuals vs Leverage plot(Figure.5) helps us to find influential cases. In the plot we have, we can see that there are not a lot of influential cases and we have a horizontal cook's distance line.

Since we see a strong linear between Age and self-rated feeling of life, considering the fact that we have multiple variables, we are using MLR technique to do the modeling.

Results

Figure.6 shows the scatter plot on the plain of personal income versus the age of the respondent, painted in light blue to dark. The light blue stands for high satisfaction and dark blue stands for not satisfied about their life.

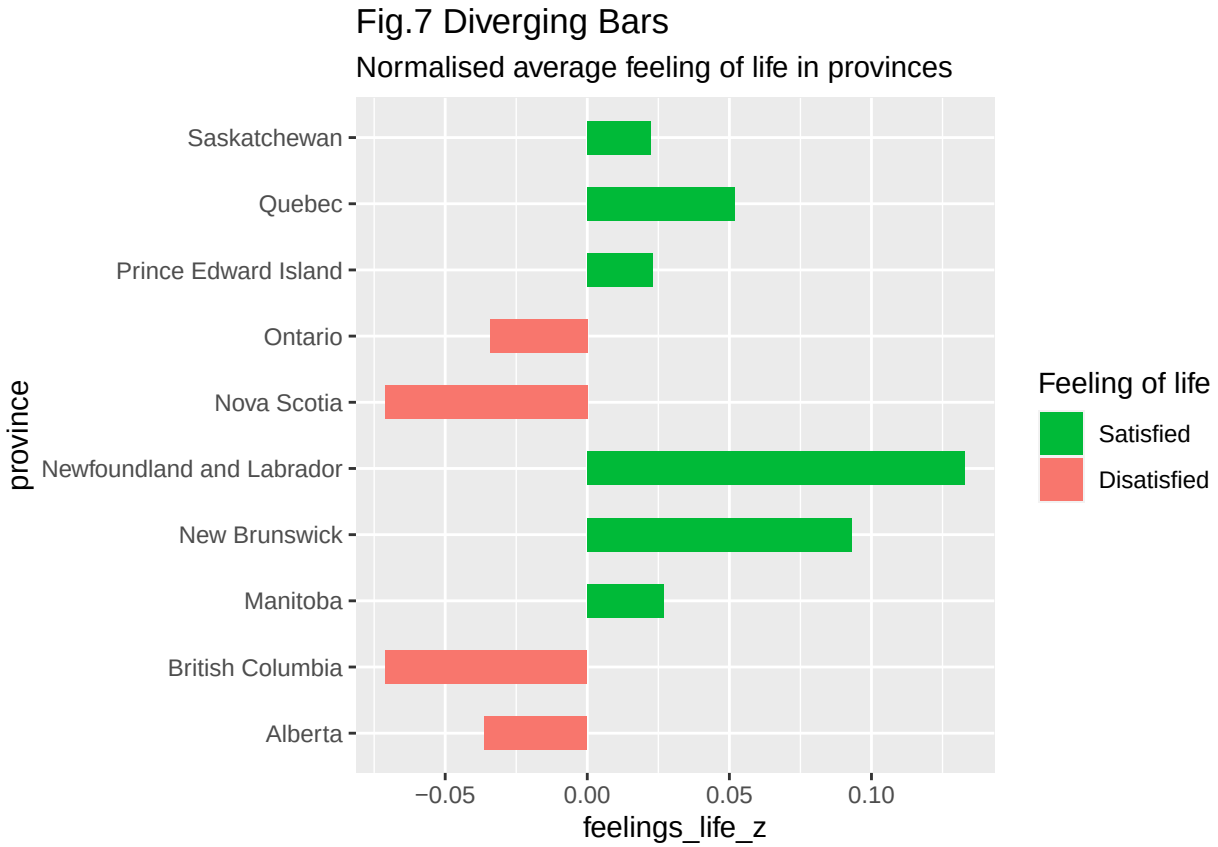


Figure.7 gives a view of normalised comparison of average feeling of life based on the province the respondent is residing in.

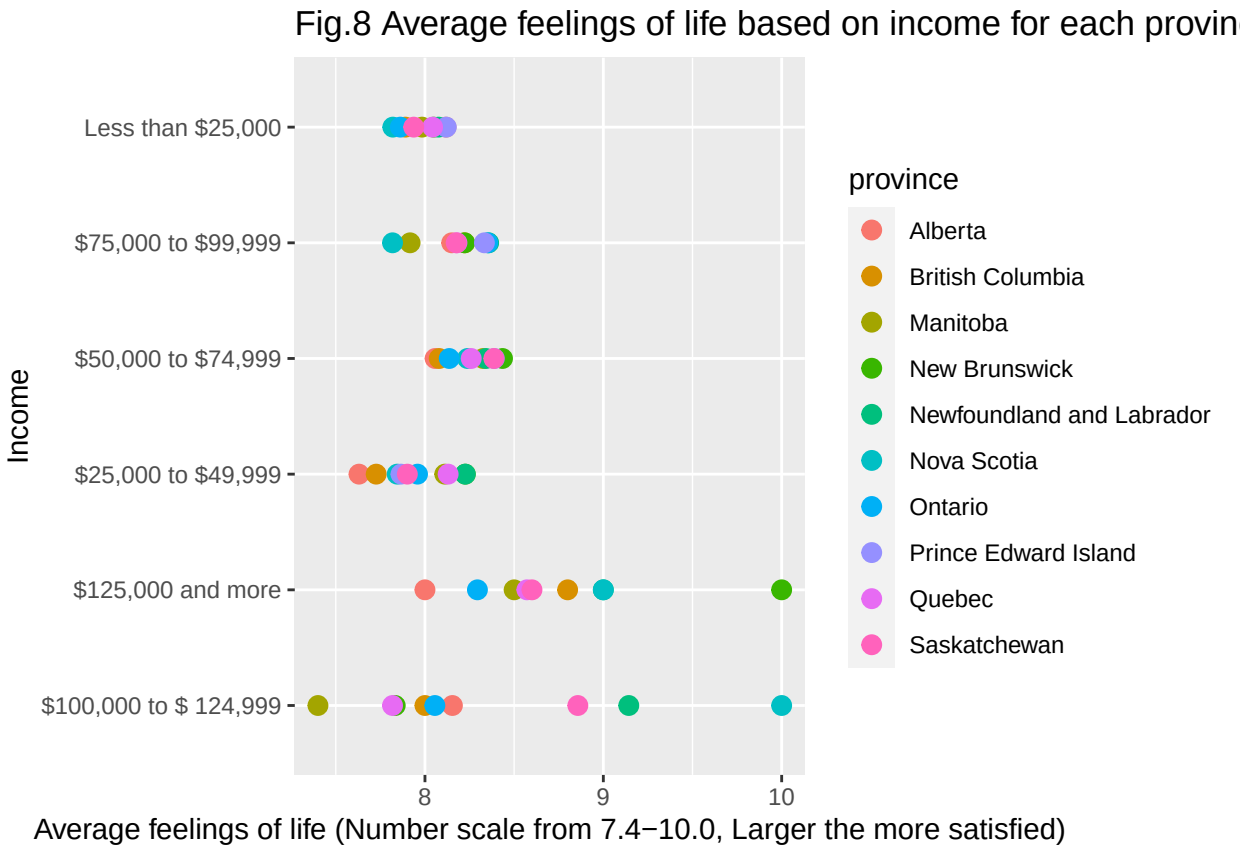


Figure 8 above shows the scatter plot of average feelings of life based on income for each province.

Table.1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.7505455	0.2446782	35.7634902	0.0000000
data$age	-0.0215585	0.0051502	-4.1859750	0.0000290
as.factor(data$income_respondent)$125,000 and more	0.3151054	0.2639234	1.1939275	0.2325725
as.factor(data$income_respondent)$25,000 to $49,999	-0.2611847	0.1776555	-1.4701756	0.1415877
as.factor(data$income_respondent)$50,000 to $74,999	0.0203756	0.1816108	0.1121937	0.9106751
as.factor(data$income_respondent)$75,000 to $99,999	0.0695580	0.1930731	0.3602679	0.7186645
as.factor(data$income_respondent)Less than $25,000	-0.3759952	0.1800622	-2.0881404	0.0368441
as.factor(data$province)British Columbia	-0.0220639	0.1010587	-0.2183277	0.8271842
as.factor(data$province)Manitoba	0.1499053	0.1197782	1.2515244	0.2108116
as.factor(data$province)New Brunswick	0.2576817	0.1267298	2.0333161	0.0420822
as.factor(data$province)Newfoundland and Labrador	0.2839240	0.1280872	2.2166464	0.0266996
as.factor(data$province)Nova Scotia	-0.0110482	0.1235105	-0.0894517	0.9287271
as.factor(data$province)Ontario	0.0306707	0.0866292	0.3540459	0.7233219
as.factor(data$province)Prince Edward Island	0.1534015	0.1502246	1.0211478	0.3072421
as.factor(data$province)Quebec	0.1885734	0.0923191	2.0426264	0.0411506
as.factor(data$province)Saskatchewan	0.0941406	0.1171630	0.8035006	0.4217300

Discussion

With the response from Table.1, fact is found that there is insufficient evidence in this sample to conclude that non-zero correlation exists. Because for those factors, p-value is greater than benchmark significance

level of 0.05.

For those factors with p-value less than 0.05, meaning the sample data provides enough evidence to build a linear relationship between them and our response variable which is self-rated feelings of life. Those factors are age (the range of age is from 15, inclusive to 35, exclusive), income (with factor less than $25000) and province (with factor 'New Brunswick' ,'Newfoundland and Labrador' and 'Quebec'). Also summary of the model gives a F-test value of 3.895 which is higher than 1.6687 which is the F critical value at freedom 15 and freedom 4281 which means this model is valid to predict the feelings of life.

The final linear equation is $Feeling.of.life = 8.75055 - 0.02156 \times X_{age} - 0.37600 \times X_{Inc<25k} + 0.25768 \times X_{NB} + 0.28392 \times X_{NL} + 0.18857 \times X_{QC} + \epsilon_i$. The estimate of intercept is 8.75055 and our age is between 15 and 35 exclusive, the intercept estimate serves the purpose of correcting output given other explanatory variables.

Slope estimates of 'New Brunswick' ,'Newfoundland and Labrador' and 'Quebec' are 0.25768, 0.28392 and 0.18857 respectively. Meaning you are expected to increase your rate of feeling by 0.25768, 0.28392 or 0.18857 if you live in 'New Brunswick' ,'Newfoundland and Labrador' and 'Quebec'. This result is consistent with what The Conference Board of Canada suggests, provinces of NB, NF and Saskatchewan have highest life satisfaction, see Figure.7. Interestingly, Saskatchewan in our model has a positive slope as well, but it is dropped due to high p-value. Quebec has a positive slope could be explained by economist Christopher Barrington-Leigh's research, suggesting that Quebec people have an increasing trend of life satisfaction due to increasing salary and more leisure time.

Slope estimate of age is -0.02156, it means you are expected to decrease your rate by 0.02156 for each additional year older. It is consistent with a popular opinion which is that happiness is "U" shaped, saying that young people and the elder have relatively higher satisfaction then the middle-age who are in the valley of the U. Since our focus is people who are between 15 and 35, it captures the 'first half of the U' and therefore, we see a decreasing trend.

The only valid slope estimate for income is -0.376(which has a P-value less than 0.05) which is for the category of income below $25000. It suggests you are expected to rate the feeling 0.376 less if your personal income is below $25000. As the figure.6 infers, this is true based on our prediction as people who has less than 25k anual income generally has a darker blue(indicating low feeling of life) than respondents with higher incomes. What's more, with a study conducted by Purdue University in 2018, if a household income is about $65000, then it reaches the threshold where more money would not necessarily increase your satisfaction(see Figure.8, which the distribution of high income average feeling of life based on province are more scattered than those with income under $75k). While it is true that the more money you have, the happier you feel before the threshold. Our result provides a rough support to this conclusion. Most households contain 2-4 persons and household income is roughly $40000-$70000 based on 25000 per person. And in our model, if personal income is below $25000, you rate feeling less, and if personal income is above $25000 (therefore the household income roughly over the 65000 threshold), you will not necessarily be happier.

It is exciting to see how this model uses real life data and draws some similar conclusions that are drawn by professional opinion, public opinion and university research. And this model empowers us to see the world in a scientific statistical way.

Some weakness exists as well, this model is still a very fundamental and robust prediction by multiple linear regression. Lots of factors are dropped due to high p-value and this model does not really function to predict one's satisfaction based on his age, income and province.

Weaknesses

The weakness of our analysis of the survey is the selection of data. All the variables and response except 'Age' are all categorical variables. Categorical variables produced a pretty poor prediction to the response especially when our response is categorical as well. Also, in figure 3, the residuals are not matched with

the dashed line of the qq plot, which suggests it is not normally distributed. So it indicates our model only explains a small portion of the sample data.

What's more, since we focus on people who are between 15 to 35, we may break the balance of sample size ratio, that is, sample size for each strata is based on population ratio, but we filter it by adding a condition that age is 15 to 35. In this process, we may break the balance of sample size ratio and get a biased sample from each strata.

Next Steps

In the future, we need to choose some better variables to predict our response. Numerical variables are more straightforward to be used. The result of a model with numerical variables can tell you if the variable you choose has correlation with your response, so you can choose to eliminate or keep it in your linear regression model. What's more, we could do a follow-up survey to confirm each numerical answer instead of ordinal answers.

Next time we may also consider a different model to predict the data. For example, by using Bayesian method, calculating the posterior distribution of Feeling of life based on the normal distributed prior,age variable, to get a more accurate prediction and wider application of the model.

References

Johnson, G. (2019, June 17). Do you live in Canada's happiest province? The Globe and Mail. Retrieved October 18, 2020, from https://www.theglobeandmail.com/life/article-do-you-live-in-canadas-happiest-province/ Life Satisfaction. (2017, April). The Conference Board of Canada. Retrieved October 18, 2020, from https://www.conferenceboard.ca/hcp/provincial/society/life-satisfaction.aspx?AspxAutoDetectCookieSupport=1 Pappas, S. (2015, November 12). Teens Are Happier Than in the Past - Why Are Adults So Miserable? Lives Science. Retrieved October 18, 2020, from https://www.livescience.com/52771-why-teens-are-happy-adults-miserable.html Welsh, J. (2011, April 19). Happiness Is U-Shaped: It Drops in Middle Age, Rises Later. Lives Science. Retrieved October 18, 2020, from https://www.livescience.com/13788-happiness-lifetime.html

Leong, M. (2019, January 14). If money doesn't buy happiness, why are we so obsessed with getting more of it? Financial Post. Retrieved October 18, 2020, from https://financialpost.com/personal-finance/material-things-dont-define-happiness-so-why-are-we-obsessed-with-money

T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.

Long JA (2020). *jtools: Analysis and Presentation of Social Scientific Data.* R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>.

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra