How the Better Life Index is affected

Yiyun Sun & Jiayue Li

Oct.19, 2020

Title of your Report

How the Better Life Index is affected

Names of Authors

Yiyun Sun & Jiayue Li

Date

Oct.19, 2020

Abstract

Nowadays, the Better Life Index for each city becomes an important factor to evaluate whether it is a livable city or not. However, the Better Life Index is considered as the experience in many aspects of the city and family. We obtained a dataset from the 2017 General Social Survey (GSS) on the family, to predict which conditions will impact the Better Life Index by using the linear regression model based on a stratified sampling method in Canada. In the results, we observe that the change of the Better Life Index can be influenced by the family's income and province, so the government and the community should pay more attention to these facts.

Introduction

We obtain a dataset that is related to the Better Life Index (feelings_life) from the Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the family, along with some correlated variables, such as hh_size, number_marriage, total_children, income_family and province. Our goal is to find how these variables impact the Better Life Index (feelings_life) in Canada. The importance of this study is we expect our linear regression model will provide the results with a comprehensive understanding of the influences on the Better Life Index, for the convenience of the government or the community to know what they can do to increase the Better Life Index for the city, and how people can improve their living experiences as well.

In the subsequent sections, we will include data, model, results, discussion, weakness and next steps respectively. In the first step, we clean the data and get the variables we need, then we use the survey design method and construct a stratification based on the province in order to build the linear regression model. Finally, we get the results based on our model and plots.

All code and data supporting this analysis is available at: https://github.com/Carriejiayue/STA304-HW2

Data

We obtain the dataset from the 2017 GSS, which is a sample survey with cross sectional design. The content of the 2017 GSS is focused on answering the number of families in Canada, the economic

conditions of these families, how their life looks in terms of different stages of families in Canada. Two primary objectives of the 2017 GSS are, to monitor changes of Canadians living conditions and well-being; to provide the information on certain current social policy issues or emerging interests. The target population is all people who are 15 years old and older in Canada, excluding all residents of the Yukon, Northwest Territories, and Nunavut; and full-time institutions, which is 30,302,287. The frame population consists of a list of telephone numbers from Statistics Canada and a list of all dwellings within the ten provinces with the number of 39,323. The sample population is 20,602 usable responses in the data.

In terms of methodology and approach, they collect the data by telephone interviews according to the list of telephone numbers they have. The respondent is randomly chosen from each household to participate. Then they divided the total ten provinces into 27 stratas to do the stratified sampling, based on 17 Census Metropolitan Areas and 10 non-CMA areas. Finally, they did the simple random sampling among each stratum.

Moreover, they get the participants' information from the telephone companies, Census of population and The Address Register. For non-responses, they will consider reweighting them by applying "Three-stage non-response adjustment", weights for responses are adjusted to represent non-responses. In the first stage, adjustments are made for non-response without any auxiliary information, this is done independently within each stratum. In the second stage, adjustments are made for non-response with auxiliary information, so they are able to predict based on auxiliary information for each household. In the third stage, adjustments are made for partial non-response with some auxiliary information, they only finish part of surveys, auxiliary information is used to model propensity to respond.

As far as the strengths of data, the 2017 GSS contains large and wide population sizes which will provide us with information from various households. They reweight the sample to represent the population, for instance, each person in the sample represents 50 persons in the population. Since the survey of some respondents is not complete, it is necessary to adjust every answer's weight. Meanwhile, this data provides lots of variables which are beneficial for us to do our study, especially, we are able to pick variables as many as we want. One drawback is that the content of the 2017 GSS states that families are becoming increasingly diverse, but the only variable in the data is the number of marriages which is insufficient to study conditions about family diversity. Meanwhile, there is too much "NA" in the data, which makes some variables useless.

The advantage of the survey is that they try their best to collect more data. According to the lists of telephone information from Statistic Canada, they tried to recontact the people who refused to answer the survey many times patiently until persuaded them to fill out. As well, the questions are comprehensive in the survey, which will help them to get more information that they want. However, in terms of limitations, they eliminate the income in the survey, but merge the income variables from previous data, which will impact the accuracy of the results. They ignore people who are not in the list, these people may don't have telephones or settle down in Canada, but their answers are still significant for the results.

Additionally, we only pick six variables to construct our own dataset. They are feelings_life, hh_size, number_marriage, total_children, income_family and province. Feelings_life is people rate their feelings of life from 0 to 10, which is similar to the Better Life Index towards our goal. Hh_size is the number of the houses in each household, since more houses will bring them higher future returns, which may impact their happiness. Number_marriage is the number of marriages for each participant. We want to see whether marriage will have influences on feelings of life. Total_children is the number of children in each household, which has always been a measure of happiness and better life. Income_family is the income of each household per year. We choose family income instead of respondent income because we

think family income influences more on how a person feels about life. Since distinct provinces will have different feelings of life due to different environments, we chose provinces as the last variable. There are still many relevant variables in the data that we didn't use, since the Better Life Index has been impacted in complex facts, we only choose that we think they are the most relative ones.

This selected data we used in this analysis contained only six variables from the original data and the last one "fpc" represents the strata for stratified sampling.

Model

In this part, we build a multiple linear regression model in R markdown to figure out the relationship between the response variables （feelings_lifeand five predictor variables (hh_size, number_marriage, total_children, income_family and province) by using a stratified sampling method. For the response variable, we treat it as the numerical variable, because we think the exact number of feelings_life represents how respondents feel and help us to figure out how feelings_life changes when the predictor variables change, even just a little change. Thus, a linear regression model is the most suitable one for this analysis.

In the first step of stratified sampling, we divide the population into ten strata depending on ten provinces and each strata size is the population of each province in 2017 since our data is from 2017. Then we construct the regression model based on the formula: feelings_life = β_0 + $\beta_1$total_children +$\beta_2$hh_size + $\beta_3$number_marriages + β_4IC1 + β_5IC2 + β_6IC3 + β_7IC4 + β_8IC5 + β_9BC + β_{10}MA + β_{11}NB + β_{12}NL + β_{13}NS + β_{14}ON + β_{15}PEI + β_{16}QU + β_{17}SA.

```
##
## Call:
## svyglm(formula = feelings_life ~ total_children + hh_size + number_marriages +
##     as.factor(income_family) + as.factor(province), design = gss.design)
##
## Survey design:
## svydesign(id = ~1, data = gss_1, strata = ~province, fpc = ~fpc)
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                                7.99025    0.06155 129.827
## total_children                             0.06544    0.01005   6.514
## hh_size                                    0.02158    0.01076   2.005
## number_marriages                           0.11784    0.02461   4.788
## as.factor(income_family)$125,000 and more  0.04647    0.03943   1.178
## as.factor(income_family)$25,000 to $49,999 -0.41876    0.04539  -9.226
## as.factor(income_family)$50,000 to $74,999 -0.25791    0.04473  -5.766
## as.factor(income_family)$75,000 to $99,999 -0.14177    0.04474  -3.169
## as.factor(income_family)Less than $25,000  -0.83762    0.05589 -14.986
## as.factor(province)British Columbia         0.03262    0.05165   0.631
## as.factor(province)Manitoba                 0.09219    0.06286   1.467
## as.factor(province)New Brunswick            0.24907    0.05957   4.181
## as.factor(province)Newfoundland and Labrador 0.23284   0.06259   3.720
## as.factor(province)Nova Scotia              0.08247    0.05994   1.376
## as.factor(province)Ontario                  0.05005    0.04558   1.098
## as.factor(province)Prince Edward Island     0.17394    0.06929   2.510
## as.factor(province)Quebec                   0.17646    0.04706   3.750
## as.factor(province)Saskatchewan             0.16973    0.06222   2.728
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## total_children                            7.51e-11 ***
## hh_size                                   0.044982 *
```

```
## number_marriages                               1.69e-06 ***
## as.factor(income_family)$125,000 and more      0.238635
## as.factor(income_family)$25,000 to $49,999      < 2e-16 ***
## as.factor(income_family)$50,000 to $74,999     8.21e-09 ***
## as.factor(income_family)$75,000 to $99,999     0.001533 **
## as.factor(income_family)Less than $25,000       < 2e-16 ***
## as.factor(province)British Columbia            0.527742
## as.factor(province)Manitoba                    0.142492
## as.factor(province)New Brunswick               2.92e-05 ***
## as.factor(province)Newfoundland and Labrador 0.000200 ***
## as.factor(province)Nova Scotia                 0.168868
## as.factor(province)Ontario                     0.272192
## as.factor(province)Prince Edward Island        0.012067 *
## as.factor(province)Quebec                      0.000177 ***
## as.factor(province)Saskatchewan                0.006381 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.59948)
##
## Number of Fisher Scoring iterations: 2
```

Table 1.1

According to the summary shown above, our primary linear regression model is feelings_life = 7.99025 + 0.06544 total_children + 0.02158 hh_size + 0.11784 number_marriages + 0.04647 IC1 - 0.41876 IC2 - 0.25791 IC3 - 0.14177 IC4 - 0.83762 IC5 + 0.03262 BC + 0.09219 MA + 0.24907 NB + 0.23284 NL + 0.08247 NS + 0.05005 ON + 0.17394 PEI + 0.17646 QU + 0.16973 SA [EQ1] which shows the relationship clearly.

Among these predictor variables X, total_children, hh_size and number_marriage are numerical variables and others are categorical variables. Five IC variables are dummy variables of income_family: when income of family is $125,000 and more, IC1 = 1, otherwise, IC1 = 0; when income of family is between $25,000 and $49,999, IC2 = 1, otherwise, IC2 = 0; similarly, IC3 means the income of family is between $50,000 and $74,999, IC4 means the income of family is between $75,000 and $99,999 and IC5 means the income of family is less than $25,000. When the income of a family is between $100,000 and $124,999, IC1 to IC5 all equal to 0.

The remaining 9 variables are dummy variables of the province: when the province is British Columbia, BC = 1, otherwise BC = 0; when the province is Manitoba, MA = 1, otherwise MA = 0; likewise, each other province variable = 1 when it is satisfied otherwise equals to 0. When the province is Alberta, all nine province variables equal to 0.

We use the specific numerical variables to demonstrate the number of houses, children, marriages for each respondent other than making them into groups, because most answers to these three variables are concentrated on several numbers, for example, most people have only 0 or 2 marriages, and fewer people have an answer greater than 2. Thus, dividing into different groups might cause the gap between each group and the difference in the number of people in each group is wide. However, for the other two categorical variables, it is better to use grouping because the happiness index of people in each province and each family with different incomes is distinct, actually.

```
##
## Call:
## svyglm(formula = feelings_life ~ total_children + hh_size + number_marriages +
##     as.factor(income_family) + as.factor(province), design = gss.design)
##
```

```
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2229  -0.6556   0.0517   0.9939   3.3233
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              7.990251   0.056584 141.211
## total_children                           0.065443   0.008979   7.289
## hh_size                                  0.021578   0.009939   2.171
## number_marriages                         0.117837   0.021178   5.564
## as.factor(income_family)$125,000 and more  0.046470   0.041786   1.112
## as.factor(income_family)$25,000 to $49,999 -0.418758   0.043772  -9.567
## as.factor(income_family)$50,000 to $74,999 -0.257913   0.044149  -5.842
## as.factor(income_family)$75,000 to $99,999 -0.141769   0.046026  -3.080
## as.factor(income_family)Less than $25,000  -0.837615   0.048346 -17.325
## as.factor(province)British Columbia       0.032615   0.045305   0.720
## as.factor(province)Manitoba               0.092190   0.067908   1.358
## as.factor(province)New Brunswick          0.249066   0.084899   2.934
## as.factor(province)Newfoundland and Labrador  0.232842   0.099722   2.335
## as.factor(province)Nova Scotia            0.082469   0.077658   1.062
## as.factor(province)Ontario                0.050050   0.037867   1.322
## as.factor(province)Prince Edward Island   0.173935   0.179315   0.970
## as.factor(province)Quebec                 0.176463   0.041239   4.279
## as.factor(province)Saskatchewan           0.169733   0.071928   2.360
##                                          Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## total_children                           3.24e-13 ***
## hh_size                                   0.02994 *
## number_marriages                         2.67e-08 ***
## as.factor(income_family)$125,000 and more  0.26610
## as.factor(income_family)$25,000 to $49,999  < 2e-16 ***
## as.factor(income_family)$50,000 to $74,999  5.24e-09 ***
## as.factor(income_family)$75,000 to $99,999  0.00207 **
## as.factor(income_family)Less than $25,000   < 2e-16 ***
## as.factor(province)British Columbia       0.47159
## as.factor(province)Manitoba               0.17461
## as.factor(province)New Brunswick          0.00335 **
## as.factor(province)Newfoundland and Labrador  0.01956 *
## as.factor(province)Nova Scotia            0.28827
## as.factor(province)Ontario                0.18627
## as.factor(province)Prince Edward Island   0.33206
## as.factor(province)Quebec                1.89e-05 ***
## as.factor(province)Saskatchewan           0.01830 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.613 on 20285 degrees of freedom
## Multiple R-squared:  0.04251,    Adjusted R-squared:  0.04129
## F-statistic: 52.98 on 17 and 20285 DF,  p-value: < 2.2e-16
```

Table 1.2

The R^2 of this model is 0.04251 which means only 4.3% variation in people's Better Life Index can be explained by the model. However, the total p-value of this model is smaller than $2.2e^{-16}$ which is small enough to represent this model is significant.

And when we look into the p-value of each variable, the p-values of IC1, BC, MA, NS, ON are about 0.24, 0.53, 0.14, 0.17, 0.27 respectively which are all greater than 0.05. This means that these five variables do not fit this model well and have very little effect on people's Better Life Index. So we can construct a new alternative model: feelings_life = 7.99025 + 0.06544 total_children + 0.02158 hh_size +

0.11784 number_marriages - 0.41876 IC2 - 0.25791 IC3 - 0.14177 IC4 - 0.83762 IC5 + 0.24907 NB + 0.23284 NL + 0.17394 PEI + 0.17646 QU + 0.16973 SA [EQ2].

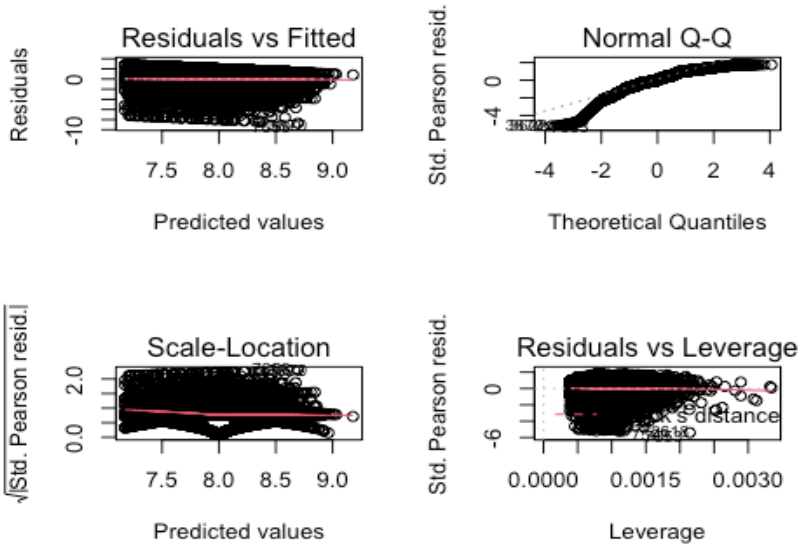

Figure 1.3

To diagnose a model, we need to see if the residuals of this model have constant variance and whether the normality of this model has been satisfied. The Residuals vs Fitted plot is the residual plot containing almost no pattern which shows the residuals of this model have constant variance and can be used to check model patterns. The Normal QQ plot represents an approximate one-to-one relationship which means the normality of this model is satisfied. And there exist many leverage points in the leverage plot as well. Overall, this model is proper and the performance is not bad.

In general, the strength of the new model is that every independent variable X has a strong influence on the response variable – feelings_life, which means this model can be used to predict for future analysis or research on the Better life index of Canadian. The weakness of this model is that the actual predictor variable of this model is only five variables, and we may neglect some more significant variables. Another thing is that this is only a linear regression model, which has some limitations compared with other models.

Results

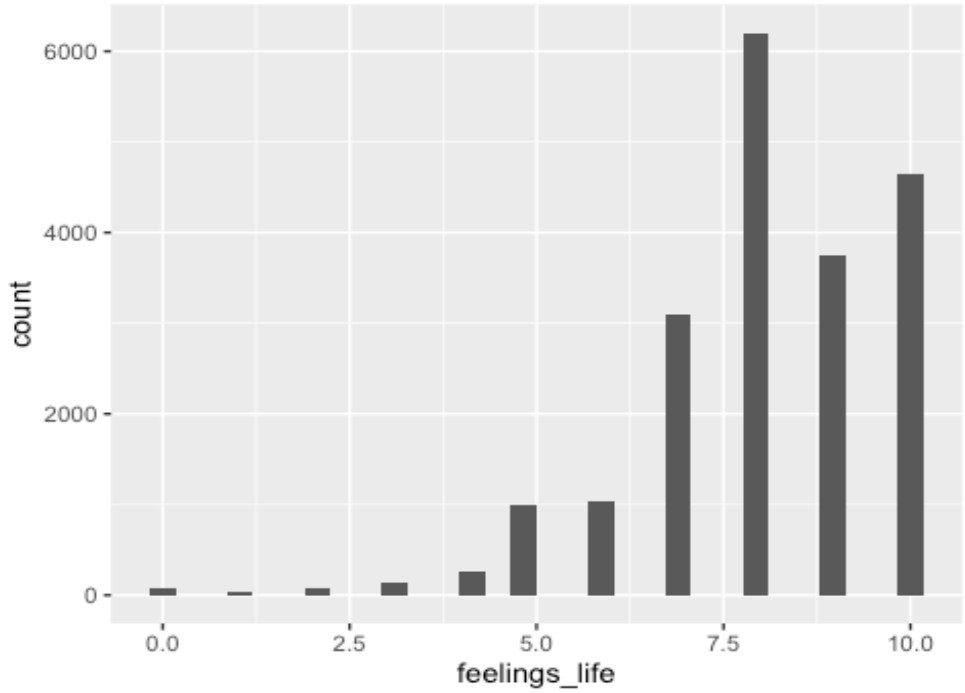

Figure 2.1

From the histogram of the feelings_life variable, it is easy to observe that the graph is a left-skewed unimodal. The Histogram graph gives a quick visual summary which represents the average feelings_life number chosen by respondents is around 7.5 with nearly 5000 respondents choosing 10 out of 10 and only less than two thousand respondents answered less than 5.

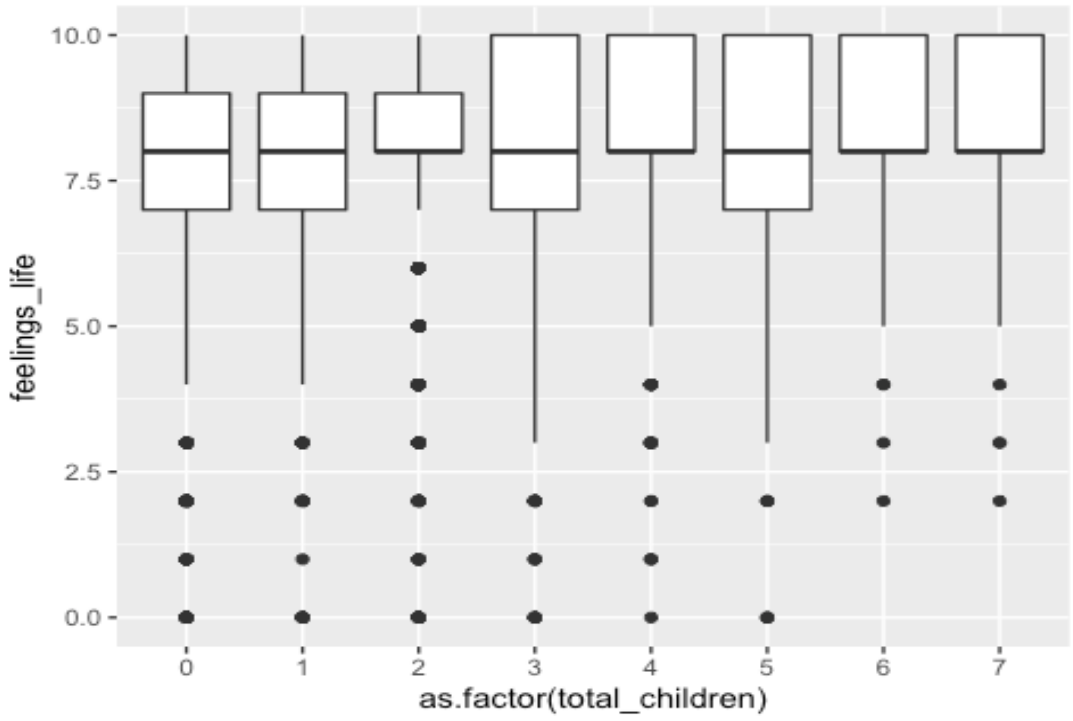

Figure 2.2

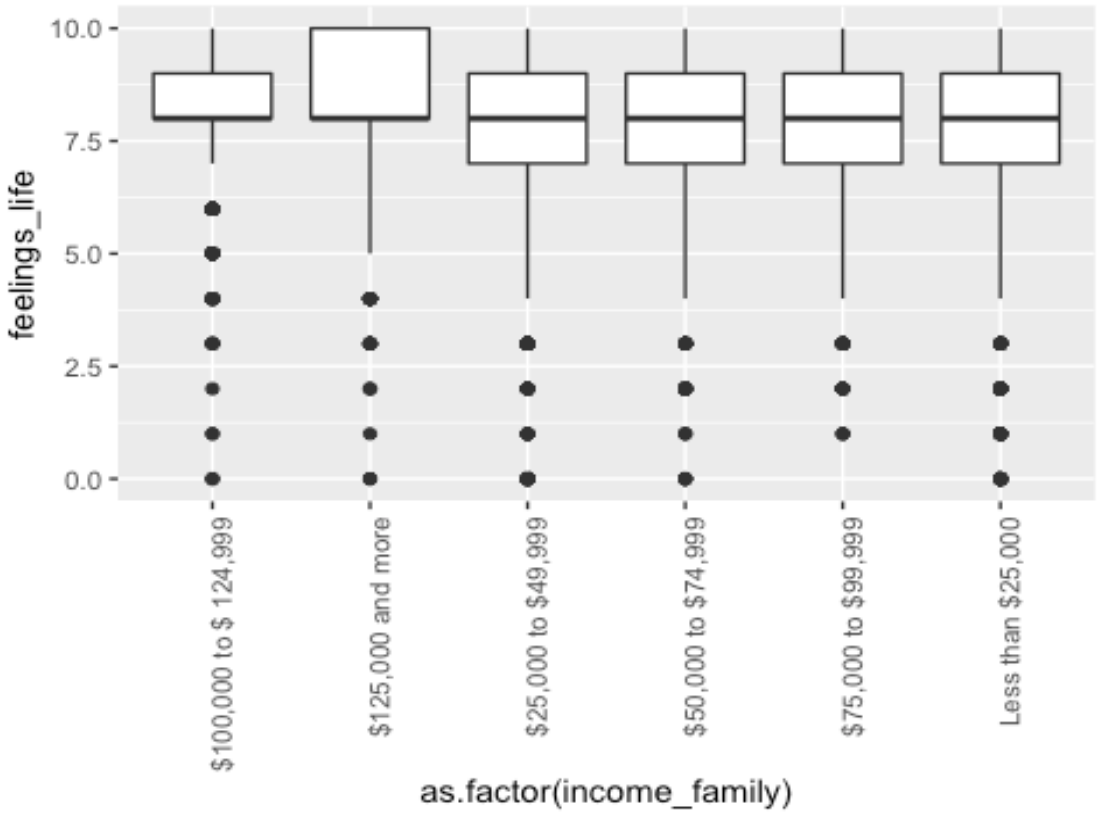

Figure 2.3

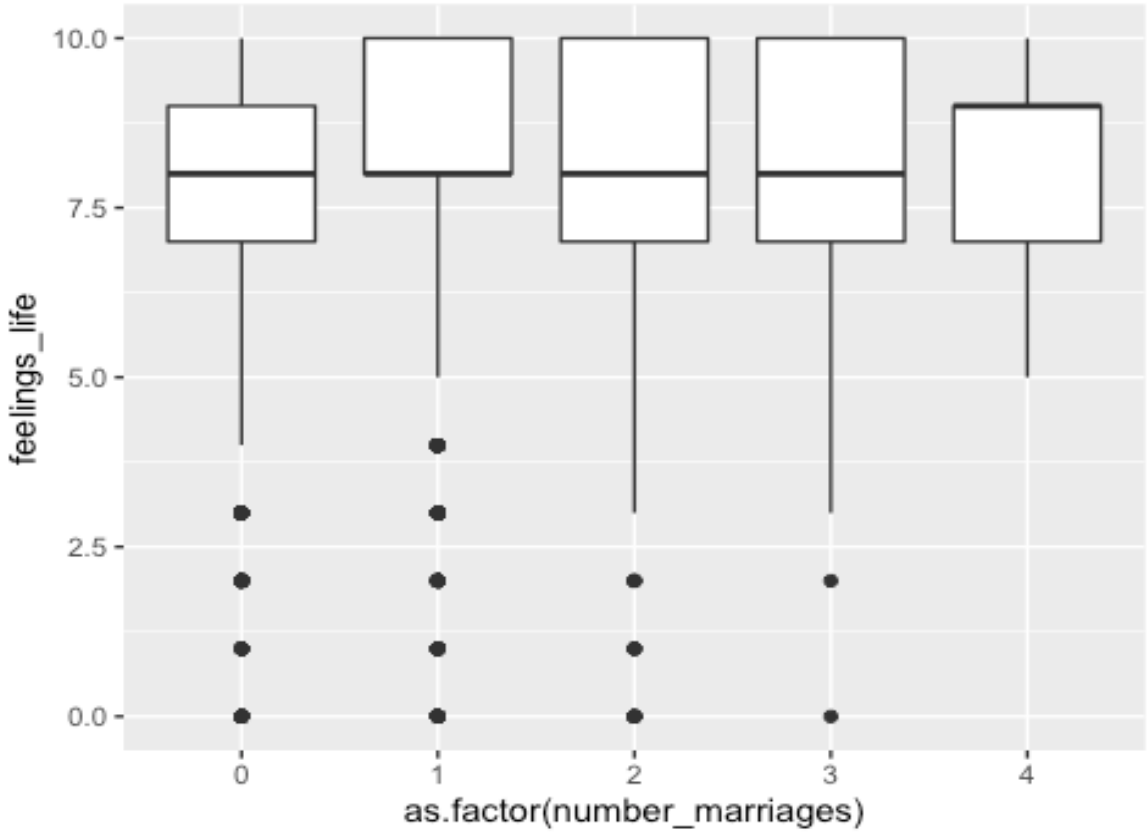

Figure 2.4

According to the p-value of each variable coefficient, the p-value of total_children, number_marriages and income_family are all small enough to conclude that these three variables contribute to the model. In order to compare the median, minimum and maximum feelings_life index of different family income, the number of marriages and the number of children, we construct three box plots which are Figure 2.2, figure 2.3, figure 2.4 respectively. From figure 2.2, it is clear to find that the median feeling_life index of the family with 0 to 7 children is the same which is 8 approximately. However, the maximum and minimum of feelings_life number change when the number of children a respondent has changed. When the total children number becomes larger than three, the range of respondent's Better Life Index is from 8 to 10 mostly. By contrast, the maximum Better Life index of respondents with less than three children is only around 8.5.

Figure 2.3 demonstrates a completely different plot. When the income of a family is less than 100,000, the minimum, median and maximum of the respondent's feelings_life index is all the same which are nearly 7, 8 and 8.8. Only the income of a family is at least 125,000, the maximum of respondent's feeling_life index is 10.

Figure 2.4 represents when the respondents have one marriage, the scope of their answer is between 8 and 10. However, the lower answer bound will become 7 if respondents have more than one marriage and the higher bound will become only 8.8 while respondents have four marriages.

Our final model is feelings_life = 7.99025 + 0.06544 total_children + 0.02158 hh_size + 0.11784 number_marriages - 0.41876 IC2 - 0.25791 IC3 - 0.14177 IC4 - 0.83762 IC5 + 0.24907 NB + 0.23284 NL + 0.17394 PEI + 0.17646 QU + 0.16973 SA [EQ2]. When the respondent living in Alberta has no children, no house and no marriage, but the family income of the respondent is between $100,000 and $124,999, the respondent's feelings-life answer is 7.99025, nearly 8. And for every one-unit increase in the number of respondent's total children, feelings_life number increases by about 0.06544 when other independent variables are fixed at a constant value. Similar with other attributes: 0.02158 hh_size and 0.11784 number_marriages. Nevertheless, the coefficient - 0.41876 means if the family income respondent is in the range from $25,000 to $49,999 instead of $100,000 to $124,999 and other conditions remain unchanged, feelings_life number decreases by 0.02158. Coefficients - 0.25791, - 0.14177, - 0.83762 have the similar meaning. The definition of next coefficient 0.24907 is if the respondent is from New Brunswick rather than Alberta and other conditions remain unchanged, feelings_life number rises by about 0.24907. Same meanings for remaining coefficients 0.23284, 0.17394, 0.17646, and 0.16973.

Discussion

From previous sections, we can conclude that the change of Better Life Index (feelings_life) can be affected by income_family and province a lot in Canada. Also, a slight impact with total_children, hh_size and number_marriages. According to Figure 1.1, we can easily observe that when a family's income per year is between $25,000 to $49,999, this has the most negative impact on the Better Life Index, which shows the income of each household between this range is not satisfied with their lives, they may don't have enough income to support their living expenses in Canada. On the other hand, Canadians who are living in New Brunswick have the most positive impacts on the Better Life Index, which may represent New Brunswick has a pleasant natural environment and characteristic cultural environment, so people are glad that they are living there.

Based on Figure 2.1, it demonstrates that most Canadian own a higher Better Life Index (feelings_life) which is 8/10. The second most Canadian have the Better Life Index with 10/10 which is pretty high. Thus, we notice that most people have great life experiences in Canada, the government and the community have done very well. Regarding Figure 2.1, we recognize that when each household has 4,6,7 children, the Better Life Index tends to be higher than others, which provides us with more children and will bring more joy to people. In accordance with Figure 2.2, it exhibits that the households with an income of $125.000 and more per each have the highest feelings_life. Obviously, wealth brings happiness

and provides better lives for people. Canadians with a single marriage have a higher Better Life Index with reference to Figure 2.3. Everyone dreams to own a happy marriage, but it doesn't mean the more marriages you have and you will be happier. In general, the first marriage will carry the most joy and expectations.

In conclusion, according to the results we get above, Canadians are able to calculate the Better Life Index (feelings_life) score by using the EQ2 we obtain. After they get the Better Life Index, they can also compare with our results to see which parts they should improve in order to achieve better living experiences. Meanwhile, the government could provide allowances to encourage people to have more children. Also, the government may consider lowering the income tax, so Canadians will get more income to improve their living experiences. In terms of community, they can construct some like "Resolve Marital Conflicts" institutions, which is for the public welfare in order to decrease the divorce rates. By doing so, Canada will gain a higher Better Life Index so that to attract more people to settle down in Canada. These are the solutions for how to increase the Better Life Index from the perspectives of the government and the community.

We obtain this dataset from the 2017 GSS data as we introduce in the Data section. First of all, we determine our objective of how the Better Life Index (feelings_life) will be affected. Secondly, we pick the variables that are correlated with our objective from the data, these variables also have less "NA" in order to get more useful data. Then, we build the model and plots to get the results. Last but not least, we analyze the results we get from the previous parts and conclude that what will have influences on the Better Life Index. As well, how people, the government and the community can improve the Better Life Index.

Weaknesses

In terms of weakness of the survey, they eliminate the income in the survey but merge the income variables from previous data, which will impact the accuracy of the results. They ignore people who are not on the list, these people may don't have telephones or settle down in Canada, but their answers are still significant for the results. Maybe in the future, the investigator can interview residential areas directly instead of telephone interviews if they have enough time and money. According to our sampling approach, even though we divide our data set into ten strata based on the ten provinces, this is still not enough compared with the 2017 GSS. We may need to get more strata like the GSS to get more precise results to improve it. For our whole analysis process, the biggest weakness is that there is not a strong linear relationship between the predictor variables and the response variables according to the diagnosis of the model. Thus, the R^2 of this model is small as well and may cause random errors in the future prediction. Although small R^2 is not a serious problem, we can consider adding more different variables in future research to increase R^2 and fix the linear relationship.

Next Steps

We aim to study the factors that affect the Better Life Index, the Better Life Index is measured based on social health, social welfare, social civilization, etc. However, we only consider the personal information which is insufficient to determine the impacts. We need to collect more data from the perspectives of society, which will improve the accuracy of our results. Furthermore, constructing more models such as the Bayesian model to analyze this data is necessary which is useful to look into the data.

References

1. Population estimates on July 1st, by age and sex. (2020, October 17). Government of Canada, Statistics Canada. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501

2. Population estimates, quarterly. (2020, October 18). Government of Canada, Statistics Canada. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901&cubeTimeFrame.startMonth=07&cubeTimeFrame.startYear=2017&cubeTimeFrame.endMonth=01&cubeTimeFrame.endYear=2018&referencePeriods=20170701%2C20180101

3. ggplot2 title : main, axis and legend titles - Easy Guides - Wiki - STHDA. (2018). STHDA. http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles

4. my.access — University of Toronto Libraries Portal. (2017). CHASS Data Center. https://login.library.utoronto.ca/index.php?url=https://sda.artsci.utoronto.ca/sdaweb/html/gss.html

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