People's feelings of life prediction using multiple linear regression models: A study on GSS (General Social Survey) 2017.

Erdong Zhang, 1004419578

Qiyun Wang, 1004006666

Yiqi Wang, 1004820964

Yue Liu, 1003937671

Date: Oct 19, 2020

Code and data supporting this analysis are available at:

https://github.com/monawang517/factors-age-children-marital-status-health-affect-people-s-feelings-of-life-in-GSS2017

##Abstract:

The pursuit of a person's life is to obtain happiness and be satisfied with their life, so increasing our overall life satisfaction is something that we are always striving for, however, there are many factors related to our perception of life so what factors affect people's feelings about their life as a whole is what we want to research in this project. Using the General Social Survey (GSS): *Families* dataset of the year 2017, we built a multiple linear regression model that can predict the score of people's feelings about their life as a whole using a given set of attributes (age, health, mental health…). We found that variables: age, total children, the age at which the first child is born, marital status, health and mental health are affected people's feeling of life. If someone is in the doldrums of his/her life period, he/she can try to change some of these attributes to change his/her current state of life and increase the score of his/her feelings about life.

##Introduction:

How happy are people today? How to measure whether they are happy? What are the factors that may lead them to be happy? are questions that we are interested in.

From the research we did, Esteban and Max[1], they concluded that factors affecting happiness include income, health and so on. We are also interested in some other parts: Are different ages matters of people to be happy? Whether having more children will make the family happier? Whether living with family or living alone affects people's feelings of life? Does living in different provinces make a difference to happiness? Are people who work fewer hours happier? Are people with higher income happier? Are healthy people happier? etc.

In this report, we will discuss factors: age, the total number of children, the age at which the first child is born, gender, living area (province), marital status, aboriginal, education level, living with family or alone, average hours of work in a week, health, mental health, income and occupation and explore their relationship with people's feeling of life(happiness).

We found that attributes: age, total children, the age at which the first child is born, marital status, health and mental health are having a positive or negative relationship with people's feeling of life.

In the small world, our result could help a person out of trouble and make some change in some aspect in his/her life to make him/her feel happier. In the big world, the government could use this model to predict people's happiness index.

The structure of this report will be:

- Abstract: the background and the reason why we choose this topic, briefly introduce the methods we use and the general conclusion we got.
- Introduction: the specific goal of our research and how to achieve it.
- Data: introduce the survey and sampling methods for how the dataset been collected, intro for the variables inside of the dataset
- Model: how to make the best fit model, how to select variables to the final model
- Results: use plots and words to describe six variables that we selected in our final model in a statistical way
- Discussion: use plots and paragraphs to describe six variables that we selected in our final model and explain their uses in real life
- Weaknesses: the problems that we meet in this report
- Next Steps: methods that we could improve in the future
- References: references for our research paper

##data:

We obtained our dataset from the Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the *families*, the dataset contains 20602 observations and 81 variables.

After reading the User Guide provided by the original research team, we know that the original data (2017 GSS) was collected from February 2nd to November 30th, 2017, is a sample survey with cross-sectional design. Its target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces (excluding Yukon, Northwest Territories and Nunavut) of Canada. The survey uses a new frame, created in 2013, that combines telephone numbers in use (both landline and cellular) with Statistics Canada's Address Register, which is a list of all dwellings within the ten provinces and collects data via telephone. In order to carry out sampling, they divided the ten provinces into strata (i.e., geographic areas), then assigned each

record in the survey frame to a stratum within its province. They are using a simple random sample without replacement sampling method next in each stratum.

All respondents in the ten provinces were rostered and interviewed by telephone. (Households without telephones were excluded from the survey population.) Survey estimates were weighted adjusted to represent all persons in the target population, including those not covered by the survey frame. For the 2017 GSS, 91.8% of the selected telephone numbers reached eligible households. To be eligible, a household had to include at least one person 15 years of age or older. During collection, households that did not meet the eligibility criteria were terminated after an initial set of questions. A respondent was then randomly selected from each household to participate in a telephone interview. Those who at first refused to participate were re-contacted up to two more times to explain the importance of the survey and to encourage their participation. For those cases where there was no one home, numerous callbacks were made.

The target sample size which is the desired number of respondents for the 2017 GSS was 20,000 while the actual number of respondents was 20,602. Finally, the overall response rate for the 2017 GSS was 52.4%.

The survey was completed by telephone and the advantage of it is the research result can be Quick collected due to the phone interviews are immediate and skilled interviewers can complete a lot of surveys in a short period of time, however, the weakness still exists, nowadays, many telephone calls are considered telemarketing, even they explained their purpose and the importance of the survey, with a long time of questioning, people will have a negative attitude and give a not so accurate response and that might influence the result of the whole research.

After we decided to make people's feeling of life as our dependent variable, we go through the 81 variables and pick out the factors we think might influence people's feeling of life (e.g. age, sex, province, education, occupation, marital status etc.). During this process, we also avoid duplicating factors such as province and region. We do not combine any factors and make it as a new variable.

```
library(tidyverse)
data <- read.csv("~/Desktop/sta304 ps2/gss.csv")
str(data)
```

##model:

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
data <- read.csv("/Users/yue/Desktop/STA304PS2/gss.csv")

#choose data
data1 <- data %>% select(2,3,5,11,12,19,22,23,28,30,38,41,42,48,49)

reg <- lm(feelings_life ~ age + total_children + age_first_child + sex + province + marital_status + aboriginal + education + living_arrangement + average_hours_worked + self_rated_health + self_rated_mental_health + income_respondent + occupation, data = data1)

broom::tidy(summary(reg))

```
We continue our study by using multiple linear regression models to predict the happiness of people. We initially select 14 candidate variables[1] from the dataset that are “age”[2], “total_children”[3], “age_first_child”[4], “sex”[5], “province”[6], “marital_status”[7], “aboriginal”[8], “education”[9], “living_arrangement”[10], “average_hours_worked”[11], “self_rated_health”[12], “self_rated_mental_health”[13], “income_respondent”[14] and “occupation”[15]. The reason we choose these 14 variables as the candidate is because they have a fairly large amount of available data to analyze in common and we believe these variables have a closer relationship with the happiness of Canadian citizens. We also avoid duplicating factors such as income_respondent and income_family. However, after running a multiple linear regression and the summary of this model, it shows that our model contains 14 independent variables which are very poor in prediction and interpretation since approximately 80% coefficients of estimators are not statistically significant. The poor performance of the initial model indicates there might be some collinearity problem and other unknown issues in the candidate variables. Hence, we decided to use a likelihood ratio test to determine the significance of a specific variable in the regression model.

#variables that could not be deleted: age, total number of children,income
```{r, echo =FALSE, message=FALSE}
#install.packages('lmtest')
library(lmtest)
reg_age_delete <- lm(feelings_life ~ total_children+age_first_child+sex+province+marital_status+aboriginal+education+living_arrangement+average_hours_worked + self_rated_health + self_rated_mental_health +income_respondent + occupation, data = data1)
```

```
reg_total_children_delete <- lm(feelings_life ~
age+age_first_child+sex+province+marital_status+aboriginal+education+living_arrangement+a
verage_hours_worked + self_rated_health + self_rated_mental_health +income_respondent +
occupation, data = data1)


reg_income_delete<- lm(feelings_life ~
age+total_children+age_first_child+sex+province+marital_status+aboriginal+education+living_
arrangement+average_hours_worked+self_rated_health+self_rated_mental_health+occupation,
data = data1)

lrtest(reg,reg_age_delete)

lrtest(reg,reg_total_children_delete)

lrtest(reg,reg_income_delete)

```

#variables that could be deleted: sex, living_arrangement, province
```{r, echo =FALSE, message=FALSE}

library(lmtest)
reg_sex_delete <- lm(feelings_life ~
age+total_children+age_first_child+province+marital_status+aboriginal+education+living_arran
gement+average_hours_worked+self_rated_health+self_rated_mental_health+income_responde
nt+occupation, data = data1)

reg_delete_living_arrangement <- lm(feelings_life ~
age+total_children+age_first_child+sex+province+marital_status+aboriginal+education+average
_hours_worked+self_rated_health+self_rated_mental_health+income_respondent+occupation,
data = data1)

reg_delete_province<- lm(feelings_life ~
age+total_children+age_first_child+sex+marital_status+aboriginal+education+average_hours_w
orked+self_rated_health+self_rated_mental_health+income_respondent+occupation, data =
data1)
```

```
lrtest(reg,reg_sex_delete)
lrtest(reg,reg_delete_living_arrangement)
lrtest(reg,reg_delete_province)
```

The likelihood ratio test is basically evaluating the goodness of the fit of two multiple linear regression models in our case. We first exclude one of the variables out of the 14 and build a new linear regression model with respect to the rest of 13 variables. Then we compare our initial model which contains 14 variables to the new model which contains 13 variables by using a likelihood ratio test. If the P-value of likelihood ratio test is smaller than 0.05 (5% significance level), it indicates that we can exclude the variable (reject the null), otherwise, the variable should be included in the linear regression (fail to reject the null). After 14 times of likelihood ratio test on each variable, the result shows that variables named "sex", "living_arrangement", and "province" could be omitted since they do not have a strong relationship with the happiness of Canadian citizens. In fact, the model does improve after we excluding the variables "sex", "living_arrangement", and "province" since the p-value of likelihood ratio test is smaller than 0.05 (significant at 5% level).

Variables that contain NA data values:age_first_child, marital_status, aboriginal,education,average_hours_worked,self health, self mental health,occupation

However, many variables (i.e. age_first_child, marital_status, aboriginal, education, average_hours_worked, self_rated_health, self_rated_mental_health and occupation) contain missing values (NA). It could not simply be deleted because NA in these variables are meaningful. For example, the NA in age_first _child is a way to show that person may not have a child. It results in the inconsistency in the quantity of data for different variables and hence unable to use likelihood ratio tests in order to do the variable selection. We decided to include those variables that contain NA also in the regression model temporarily, and named the model "reg_finaltem".

To conclude the above info, the temporary best fit model after ratio likelihood test:

```{r echo =FALSE, message=FALSE}
library(lmtest)
reg <- lm(feelings_life ~
age+total_children+age_first_child+sex+province+marital_status+aboriginal+education+living_arrangement+average_hours_worked+self_rated_health+self_rated_mental_health+income_respondent+occupation, data = data1)

reg_finaltem<-lm(feelings_life ~
age+total_children+age_first_child+marital_status+aboriginal+education+average_hours_worked+self_rated_health+self_rated_mental_health+occupation, data = data1)
```

```
lrtest(reg, reg_finaltem)
broom::tidy(summary(reg_finaltem))
```

To find out whether the variables which contain NA have close linear relationships to feelings of life, we decide to do two steps. Firstly, we check the summary table of "reg_finaltem", pull out the variables that have p-values greater than 0.05. Secondly, graph and analyze each of them separately to double-check whether they should be deleted.

#step1: see the summary of reg_finaltem and select the p-value greater than 0.05's variables
```{r, echo =FALSE, message=FALSE}
broom::tidy(summary(reg_finaltem))

reg_finaltem<-lm(feelings_life ~
age+total_children+age_first_child+marital_status+aboriginal+education+average_hours_worke
d+self_rated_health+self_rated_mental_health+occupation, data = data1)
broom::tidy(summary(reg_finaltem))
```
secondly, graph and analyse each of the variables that have p-value larger than 0.05 in reg_finaltem separately.

#graph of aboriginal
```{r, echo =FALSE, message=FALSE, warning = FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, color = `aboriginal`)) + geom_histogram(binwidth =
0.7) + labs(title = "different feelings of life outcome of whether aboriginal ")
aboriginal_model <- lm(data$feelings_life~data1$aboriginal)
broom::tidy(summary(aboriginal_model))
```
We were predicted that the aboriginal people have less happiness index than the people who are not aboriginal. The histogram indicated that the non-aboriginal people have a larger amount of the Happiness index. This shows that aboriginal people are not as happy as non-aboriginal people. However, in the summary of the regression line, the P-value of both aboriginal-No and aboriginal-Yes are larger than 0.05, which suggests we do not include the aboriginal variable into the model. This might because there are potential lurking variables to make them have larger p-value.

#graph of education
```{r, echo =FALSE, message=FALSE}
```

```
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, fill = education)) + geom_bar() + labs(title =
"different feelings life outcome of different level of education")
education_model <- lm(data1$feelings_life ~ data1$education)
broom::tidy(summary(education_model))
```

We want to find out whether different levels of education have a linear relationship with the feelings of life index. Firstly, we drew a bar plot by these two variables, from the graph we can see that each level of education is having approximately the same proportion in each feeling of life index. So we initially concluded that it should not have much influence on people's feelings of life index. Next, we offered a linear regression model of these two variables and looked at their summary result. Although the p-value of the total model is less than 0.05, nearly half of the variables' p-value in the data is larger than 0.05. In order to find the best-fitted model, we decided to drop the variable "education" from our final dataset.

#graph of average hours worked

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, fill = `average_hours_worked`)) + geom_bar() +
labs(title = "different feelings life outcome of different average hours worked")
```

We want to find out whether different average hours worked to have a linear relationship with the feelings of life index. Firstly, we drew a barplot by these two variables, from the graph we can see that each type of average hours worked are having approximately the same proportion in each feeling of life index, and with the large amount of "NA", in order to find the best-fitted model, we decided to drop variable "average_hours_worked" from our final dataset.

From the summary table of “reg_finaltem”, we can find out that coefficients of the variable “aboriginal”, “education” and “average_hours_worked” are not statistically significant at 5% significance level. Also, we found that in the graphs of “aboriginal”, “education” and “average_hours_worked” also shows that these three are not connected to the feelings of life. Therefore, we decide to remove them from the temporarily best fit model and conclude our final multiple linear regression model with the following formula:

$$Feelings\_life = \beta_0 + \beta_1 age + \beta_2 total\_children + \beta_3 age\_first\_child + \beta_4\, marital\_status\ Living\ Common\text{-}Law + \beta_5\, marital\_status\ Married + \ldots + \beta_8\, marital\_status\ Widowed + \beta_9\, self\_rated\_health\ Excellent + \ldots + \beta_{13}\, self\_rated\_health\ Poor + \beta_{14}\, self\_rated\_mental\_health\ Excellent + \ldots + \beta_{18}\, Self\_rated\_mental\_health\ Poor.$$

In other words, our model reveals that happiness is associated with ages, the total number of children the one has, the age of the one's first child, the marital status of the one, the physical and mental health rate of the one on average. We use "age" and "age_first_child" as a numerical variable rather than age-groups because by doing in this way, it would be more clearly to infer that a one-year increase in age of the one or his/her first child is associated with a specific level of changing in happiness on average holding everything else constant. Changing "age" and "age_first_child" into age-groups is nothing more than turning the change of a specific year into the average difference between two age-groups. In addition, this model is a built-in RStudio. The Variables "age", "total_children", and "age_first_child" are numerical variables and the rest are categorical variables. In specific, a variable named "marital_status" has 6 categorical variables, the reference categorical variable in the regression is "Divorced". Variables named "self_rated_health" and "self_rated_mental_health" both have 6 categorical variables as well, and their reference categorical variable in the regression is "Don't know". The categorical variable will become 1 if one belongs to that category, otherwise, the categorical variable is 0. The coefficients of each variable are shown in the table below.

```*{r, echo =FALSE, message=FALSE}*

*reg_final<-lm(feelings_life ~
age+total_children+age_first_child+marital_status+self_rated_health+self_rated_mental_health, data = data1)*

*broom::tidy(summary(reg_final))*

```

Diagnostic test

```{r, echo =FALSE, message=FALSE}

reg_final<-lm(feelings_life ~
age+total_children+age_first_child+marital_status+self_rated_health+self_rated_mental_health, data = data1)

par(mfrow=c(2,2))

plot(reg_final,1)

plot(reg_final,2)

plot(reg_final,3)

plot(reg_final,5)

```

We have also done the diagnostic test to test whether the final model is good.
1. Residuals vs. Fitted plot

This plot tells us if residuals have non-linear patterns. From our plot, we could find many equally spread residuals around a horizontal line without distinct patterns, that means our residuals have a non-linear relationship and our model is appropriate.

2. Normal Q-Q plot

This plot tells us whether residuals are normally distributed. From our plot, we can see we have a heavy tail Normal Q-Q plot. This plot shows most of the residuals are normally distributed, but not all of the residuals are lined well on the straight dashed line which may be caused by some problems and weakness of the original dataset that we used.

3. Scale-Location

This scale-location plot shows that the residuals begin to spread wider along the x-axis as x increases. This might be because, in our original dataset, the data is not well-defined so that the spread of the residuals could not be described.

4. Residuals vs. Leverage

This plot helps us to find influential parameters. Not all outliers are significant in linear regression analysis (whatever outliers mean). Even if the data has extreme values, they may not affect the determination of the regression line. In other words, if we include it in the analysis or exclude it from the study, the results will not be much different. The graph's rightmost point is on the line of mean 0, so this is a good leverage point, which means the Y-value of the point closely follows the pattern set by the other points and this point is not an outlier. This graph shows all points have low Cook’s distance scores, which means the parameters whose deletion from the dataset has a small effect on the parameter estimates.

The model is strong in making inferences compared to previously proposed models because the p-values of almost all coefficients of estimators are statistically significant. However, a better inferential model accompanied by inaccurate predictions. Since the adjusted R-squared is 0.2858, which means approximately 29% variation of feelings_life (Happiness) is explained by the variation of independent variables in our regression model. In addition, the mean square error is increased compared to the original model and the temporarily best fit model selected by the likelihood ratio test. (The MSE for the initial model, likelihood selected model and the final model is listed below respectively).

```{r echo =FALSE, message=FALSE}

mean(reg$residuals^2)

mean(reg_finalitem$residuals^2)

mean(reg_final$residuals^2)

```

Usually, model convergence requires iterations, and the iteration frequency is directly proportional to the accurate. Based on the knowledge we acquired so far, we are inadequate to perform the iterative process on our linear model. However, the major determinants of happiness in our model are mostly categorical variables and combine with poor performance in predicting as we discussed above. We are unconfident to propose that the model is convergent. Overall, our final model is sufficient in making inferences about the happiness of people but precision needs to be improved.

---

[1] Explanation of the variables refers to footnotes.

[2] Age in numerical value, not age groups.

[3] The total number of children the respondent has.

[4] The age of the first child in the sampling year (2017).

[5] Male or female.

[6] Ten provinces of Canada.

[7] living common-law, married, separated, single never married, widowed, divorced.

[8] Yes or No.

[9] Less than high school diploma and its equivalent, High school diploma and its equivalent, College certificate and its equivalent, University certificate or diploma below the bachelor's level, Bachelor's degree, University certificate, Diploma or degree above bachelor's level, Trade certificate or diploma.

[10] Alone, Spouse only, Spouse and single child under 25 years of age, Spouse and single child 25 years

of age or older, Spouse and other, Spouse and non-single child(ren), No spouse and non-single child(ren), No spouse and single child under 25 years of age, No spouse and single child 25 years of age or older, Living with one parent, Living with two parents, Other living arrangement.

[11] 0.1 to 29.9 hours, 30.0 to 40.0 hours, 40.1 to 50.0 hours, 50.1 hours and more, Don't know.

[12] Excellent, Very good, Good, Fair, Poor, Don't know.

[13] Excellent, Very good, Good, Fair, Poor, Don't know.

[14] Less than $25,000, $25,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $124,999, $125,000 and more.

[15] Health Occupations, Management occupations, Natural and applied sciences and related occupations, Natural resources, agriculture and related production occupations, Occupations in art, culture, recreation and sport. Occupations in education, law and social community. Occupations in manufacturing and utilities, Sales and service occupations, Trades, transport and equipment operators and related occupations, Business, finance, and administration occupations, Uncodable.

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
data <- read.csv("/Users/yue/Desktop/STA304PS2/gss.csv")

#choose variables that we think are related to feelings_of_life
data1 <- data %>% select(2,3,5,11,12,19,22,23,28,30,38, 41, 42, 48, 49)

#install.packages('lmtest')
reg <- lm(feelings_life ~ age + total_children + age_first_child + sex + province +
marital_status + aboriginal +  education + living_arrangement  +average_hours_worked +
self_rated_health + self_rated_mental_health + income_respondent + occupation, data = data1)
broom::tidy(summary(reg))
```

#graph of sex

```{r, echo =FALSE, message=FALSE, warning = FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, color = `sex`)) + geom_histogram(binwidth = 0.5) +
labs(title="Happiness Index in gender")
```

From the histogram, we could see in every score of the happiness index, the red part of the diagram (female) and the blue part of (male) almost have the same percentage which means that there is no gender difference in the Happiness index. Also, from the ratio likelihood test, the p-value of the original model and model that deleted sex variable is less than 0.05, which show that the model without the sex variable is better.

#graph of living_arrangement

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, fill = `living_arrangement`)) + geom_bar() + labs(title
= "different feelings life outcome of different type of living arrangement")
living_model <- lm(data1$feelings_life ~ data1$living_arrangement)
summary(living_model)
```

We want to find out whether different types of the living arrangement have a linear relationship with the feelings of life index. Firstly, we drew a bar plot by these two variables, from the graph we can see that each type of living arrangement is having approximately the same proportion in each feeling of life index. So we initially concluded that it should not have much influence on people's feelings of life index. Next, we offered a linear regression model of these two variables and looked at their summary result. Although the p-value of the total model is less than 0.05, many of the variables' p-value in the data is larger than 0.05. In order to find the best-fitted model, we decided to drop the variable "living_arrangement" from our final dataset.

#graph of province

```{r, echo =FALSE, message=FALSE, warning=FALSE}
library(tidyverse)
ggplot(data=data1, aes(y =`feelings_life`))+geom_bar() + geom_bar(aes(fill = `province`)) +
labs(title="Happiness Index in different province")

province_model <- lm(data1$feelings_life~data1$province)
broom::tidy(summary(province_model))
```

The diagram of the province shows that in the highest happiness index, province Ontario and Quebec occupied the larger amount of the bin. However, other provinces are kindly the same. Secondly, from the p-value, we could see that most of the provinces' p-value is larger than 0.05 which shows that we fail to reject that there is no connection between feelings of life and the province. There might be some potential lurking variables which affect this p-value.

#graph of education

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, fill = education)) + geom_bar() + labs(title =
"different feelings life outcome of different level of education")
education_model <- lm(data1$feelings_life ~ data1$education)
broom::tidy(summary(education_model))
```

We want to find out whether different levels of education have a linear relationship with the feelings of life index. Firstly, we drew a bar plot by these two variables, from the graph we can see that each level of education is having approximately the same proportion in each feeling of life index. So we initially concluded that it should not have much influence on people's feelings of life index. Next, we offered a linear regression model of these two variables and looked at their summary result. Although the p-value of the total model is less than 0.05, nearly half of the

variables' p-value in the data is larger than 0.05. In order to find the best-fitted model, we decided to drop the variable "education" from our final dataset.

#graph of average hours worked

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, fill = `average_hours_worked`)) + geom_bar() +
labs(title = "different feelings life outcome of different average hours worked")
```

We want to find out whether different average hours worked to have a linear relationship with the feelings of life index. Firstly, we drew a barplot by these two variables, from the graph we can see that each type of average hours worked are having approximately the same proportion in each feeling of life index, and with the large amount of "NA", in order to find the best-fitted model, we decided to drop variable "average_hours_worked" from our final dataset.

#graph of partner main activity

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, fill = `partner_main_activity`)) + geom_bar() +
labs(title = "different feelings life outcome of different type of partner main activity")
partner_model <- lm(data1$feelings_life ~ data1$partner_main_activity)
broom::tidy(summary(partner_model))
```

We want to find out whether different types of partner main activity have a linear relationship with the feelings of life index. Firstly, we drew a bar plot by these two variables, from the graph we can see that each type of partner's main activity is having approximately the same proportion in each feeling of life index. So we initially concluded that it should not have much influence on people's feelings of life index. Next, we offered a linear regression model of these two variables and looked at their summary result. Although the p-value of the total model is less than 0.05, many of the variables' p-value in the data is larger than 0.05, and with the large amount of "NA", in order to find the best-fitted model, we decided to drop variable "partner_main_activity" from our final dataset.

#graph of aboriginal

```{r, echo =FALSE, message=FALSE, warning=FALSE}
library(tidyverse)
ggplot(data = data1, aes(x=`feelings_life`, color = `aboriginal`)) + geom_histogram(binwidth =
0.7) + labs(title = "different feelings of life outcome of whether aboriginal ")
aboriginal_model <- lm(data$feelings_life~data1$aboriginal)
broom::tidy(summary(aboriginal_model))
```

```
We were predicted that the aboriginal people have less happiness index than the people who are not aboriginal. The histogram indicated that the non-aboriginal people have a larger amount of the Happiness index. This shows that aboriginal people are not as happy as non-aboriginal people. However, in the summary of the regression line, the P-value of both aboriginal-No and aboriginal-Yes are larger than 0.05, which suggests we do not include the aboriginal variable into the model. This might because there are potential lurking variables to make them have larger p-value.

## ##Result:

By looking at the final linear regression model, we find out that as "age, total_children, age_first_child, marital_status, self_rated_health, and self_rated_mental_health" change, "feelings_life" will be different. For example, Figure 1 is a scatterplot of the relationship between age and feelings_life, which shows the happiness of people aged 40 is from 2 to 10. Hence, although a group of people have the same age, other factors affect their feeling about life, which can be validated by our model result. By looking at Figure 2, the average feelings_life per total_children, we can tell that happiness will slightly increase as the number of total children increases. Figure 3 is the relationship between age_first_child and feelings_life, showing the age of people the first child will influence happiness. By looking at Figure 4, different marital statuses in feelings_life, we can tell married people have a large proportion of feelings about life. Figures 5 and 6 show that the happiness index increases with the improvement of the body and mental health rate.

By combining our model result, we can conclude that people's age, the number of their total children, the age of their first child, their marital status, self-rated health, and self-rated mental health will accept people's feelings about life.

### #graph of age

```{r, echo =FALSE, message=FALSE, include=F}
library(tidyverse)
ggplot(data, aes(age, feelings_life, color = feelings_life)) + geom_point() + labs(title = "Figure 1:age vs feelings_of_life")
age_model <- lm(data$feelings_life ~ data$age)
broom::tidy(summary(age_model))
```

We want to explore whether the age of respondents affect feelings about life or not. First, we drew the two variables into a scatter plot; as shown in the graph, age and feelings_life have no apparent correlation. Regardless of age, respondents had a happiness index of approximately seven and above. Next, we offered a linear model of these two variables. It shows that the p-value is less than 0.05, which is statistically significant and indicates strong evidence for the
```

assumption. In other words, the age of respondents affects their feelings about life. Moreover, the happiness index is relatively high when respondents are between 26 to 75 years old.

#graph of total children

```{r, echo =FALSE, message=FALSE, include=F}
library(tidyverse)
plot(feelings_life ~ as.factor(total_children), data = data, main = "Figure 2: total_children vs Feelings of life", xlab = "total_children")
total_children_model <- lm(data1$feelings_life~data1$total_children)
broom::tidy(summary(total_children_model))
```

As we can see, these six box plot shows mean feelings_life is 8. When respondents have two children and more, these boxplots are right-skewed, which means that the respondents' feelings about life will be higher with more children. Moreover, the linear model shows the p-value is less than 0.05, which is statistically significant and indicates strong evidence for the hypothesis. Thus, the number of total children influences respondents' feelings about life.

#graph of age first child

```{r, echo =FALSE, message=FALSE, include=F}
library(tidyverse)
ggplot(data, aes(age_first_child, feelings_life, color=feelings_life)) + geom_point() + labs(title = "Figure 3: age of first child vs feelings of life")
age_first_child_model <- lm(data1$feelings_life ~ data1$age_first_child)
broom::tidy(summary(age_first_child_model))
```

We want to explore whether the age of respondents' first child (age_first_child) affects feelings about life (feelings_life) or not. First, we drew the two variables into a scatter plot; as shown in the graph, age_first_child and feelings_life have no apparent relationship. Regardless of the age of the first child, respondents had a happiness index of 5 and above. Next, we offered a linear model of these two variables. It shows that the p-value is small, which is 0.03278, so we did not reject our conjecture. In other words, the age of respondents' first child does affect their feelings about life.

#graph of mariginal status

```{r, echo =FALSE, message=FALSE}
library(tidyverse)
ggplot(data=data1, aes(y =`feelings_life`))+geom_bar() + geom_bar(aes(fill = `marital_status`)) + labs(title="Figure 4: marriage vs Feelings of life")
marital_status_model <- lm(data1$feelings_life~data1$marital_status)
broom::tidy(summary(marital_status_model))
```

```
We are interested in whether there is a connection between marital status and feelings of life. From the plot, the married people have occupied the larger part of the high score of people's feelings of life index. Also, from the p-value table, it shows that almost every status of marriages has p-value less than 0.05 which shows that we reject the null hypothesis test about there is no connection between marital status and feelings of life.

**#graph of self_rated_health and self_rated_mental_health**

```{r, echo =FALSE, message=FALSE}
plot(feelings_life ~ as.factor(self_rated_health), data = data1, main = "Figure 5: health vs feelings of life", xlab = "self_rated_health")
plot(feelings_life ~ as.factor(self_rated_mental_health), data = data1, main = "Figure 6: mental_health vs feelings of life", xlab = "self_rated_mental_health")
self_rated_health_model <- lm(feelings_life ~ self_rated_health, data = data1)
self_rated__mental_health_model <- lm(feelings_life ~ self_rated_mental_health, data = data1)
```

From the plot, we can see that the respondents whose body health and mental health are at excellent levels tend to be the happiest group on average. In addition, both plots reveal that the happiness index decreases as the body and mental health rate level lower down. Therefore, the variables "self_rated_health" and "self_rated_mental_health" are good indicators of happiness.

##Discussion：

Our dataset is from the Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the families. The data was collected by phone sampling. The bias of phone sampling is that the sampling did not include the family that did not have phones and people who did not answer the phone. Mancini said that phone surveys could be considered market fraud so that people do not truly trust phone surveys so that people could not answer the survey questions positively. The original data contains bias and we first did some research and selected 14 variables that might be connected with feelings of life manually. We avoided the variables that have the same topic. Secondly, we use a likelihood ratio test to compare each of the variables that we have selected with the original model and we deleted sex, living_arrangement, province, and we kept age, the total number of children, income, and the NA terms we used two steps to test whether it is a good variable to keep in the model. Thirdly, we ran the codes in r to see the summary of the model we have in the last step and take out the p-value >0.05's variables and plot each of them to adjust the model.
```

We will explore relationships with happiness through six aspects: age, the total number of children, age of the first child, marital status, physical and mental health. The following six graphics can explain their correlation:

```{r, echo =FALSE, message=FALSE, include=F}
library(tidyverse)
ggplot(data, aes(age, feelings_life, color = feelings_life)) + geom_point() + labs(title = "Figure 1")
plot(feelings_life ~ as.factor(total_children), data = data, main = "Figure 2", xlab = "total_children")
ggplot(data, aes(age_first_child, feelings_life, color=feelings_life)) + geom_point() + labs(title = "Figure 3")
ggplot(data=data1, aes(y =`feelings_life`))+geom_bar() + geom_bar(aes(fill = `marital_status`)) + labs(title="Figure 4")
plot(feelings_life ~ as.factor(self_rated_health), data = data, main = "Figure 5", xlab = "self_rated_health")
plot(feelings_life ~ as.factor(self_rated_mental_health), data = data, main = "Figure 6", xlab = "self_rated_mental_health")
```

As we can see, Figure 1 shows the relationship between the age of respondents and feelings about life. Before the age of 27, most people's happiness index is relatively high, and no one's happiness index is 0, which means that before the age of 27, people feel life is happy, and there are no big worries. After the age of 27, people's concerns may increase due to life pressure, family pressure, work pressure, etc. gradually, some people's happiness index is 0. But if people can handle the problems well, they can improve their feelings about life.

If you want to know how many children will lead parents to be happier, please see Figure 2. It shows that no matter how many children they have, people's average happiness index is 8. We recommend that people have two children as the most appropriate because the lowest happiness index of those with two children is 7. The reason may be that two children can take care of each other, so parents do not need to worry too much about children's life.

Figure 3 can be used as a reference for people who have children or plan to have children. As the children get older, the feelings about life of the parents' increases. Because before the child reaches adulthood, parents need to consider the child's diet, daily life, and health. But after children reach adulthood, they need to be independent, take care of themselves, and be financially independent. Parents don't have to worry about it, and then their feelings about life will increase.

It is often said that marriage is the grave of love and that falling in love will restrict people’s freedom. But based on Figure 4, we found some suggestions for people who are afraid of falling in love or getting married. According to the graph, people who get married and live by common law are happier than people who never married. This might because the feeling of

accompaniment could make people feel happy. And marriage could make you feel not lonely. So if you want to be happier, you can find love bravely and get married boldly.

According to figure 5 and figure 6, people who do not have health problems either in the physical part or in the mental part are happier. If you are sick, you might have to go to the hospital and cure yourself. If your body is not in a healthy state, diseases will make you feel uncomfortable and you will not feel happy. The cost of going to the hospital may be unaffordable, which can also cause unhappiness. Mental health matters. People who have stressful lives or have depression will not be happy. People should really pay attention to the mental health problem because you will be unhappy.

##Weaknesses：

· Firstly, the original dataset itself contains problems with the collecting ways. In this 2017 GSS model, they used telephone surveys. Telephone surveys could be considered as marketing fraud so people may not respond positively to questions or even hang up the phone. Also, this survey could not have responses from the family who do not have phones. The dataset itself contains weakness.

· Secondly, the models we made have a small R square which is because the dataset itself has the collecting problem. R square gives us a percentage of variation in y's explained regression line. It is bad to have a small R square.

· Thirdly, the final model we made has a larger mean square error. And according to the R document that Daniel wrote, the mean square error measures the average of the squared of the errors. Low values of mean square error are a better fit model.

· Fourthly, we select the variable that we think matters to the happiness index which is a subjective choice without the scientific support.

Fifthly, we use ratio likelihood to test the model differences but the test is limited to the variables without NA, but in our variables, some NA have special meaning. For example, due to the target population of our sampling being people whose age are equal to or older than 15, the NA of age_first_child is because they are too young to have a child. We could not just simply remove NA that have special meanings. We did not find better ways to solve this problem.

##Next Steps：

- we could learn better methods of making models to produce a better fit of the model
- we could improve the methods of collecting data from the survey
- we could learn how to deal with NA that have special meanings.

##References:

Alexander, R & Caetano, S (October 7, 2020). “gss_clean.R”. Retrieved from https://q.utoronto.ca/courses/184060/files/9422740?module_item_id=1867317

Beaupré, P. (April 2020). General Social Survey Cycle 31: Families Public Use Microdata File Documentation and User’s Guide, (Issue no. 2019001). Retrieved October 18, 2020, from Chass https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/index.htm (need UTORid)

Daniel, L. Compute Model Quality. Retrieved October 17, 2020, from https://www.rdocumentation.org/packages/sjstats/versions/0.17.4/topics/cv

Esteban Ortiz-Ospina (2013) - "Happiness and Life Satisfaction". Published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/happiness-and-life-satisfaction

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

Lumley T (2004). “Analysis of Complex Survey Samples.” *Journal of Statistical Software*, **9**(1), 1-19. R package version 2.2.

Lumley T (2010). *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley and Sons.

Kassambara, Visitor, & Mann, T. (2018, March 11). Linear Regression Assumptions and Diagnostics in R: Essentials. Retrieved October 20, 2020, from http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

Mancini, C. (May 4, 2020). Advantages and disadvantages of a phone survey. Retrieved October 17, 2020, from https://www.idsurvey.com/en/advantages-and-disadvantages-of-phone-survey/

R Core Team(2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Zeileis A, Hothorn T (2002). “Diagnostic Checking in Regression Relationships.” *R News*, **2**(3), 7–10. https://CRAN.R-project.org/doc/Rnews/.