Connection between feeling of life and age, children, income:GSS 2017

Xiaoran Zhang; Haoji Ye Weichao Wu;Yaozhong Zhang

19/10/2020

#data: 'Code supporting this analysis is available at: https://github.com/Weichao-Wu/STA304-project-2'

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.2     ✓ dplyr   1.0.0
## ✓ tidyr   1.1.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

#author: Xiaoran Zhang 1004115744; Haoji Ye 1004371967; Weichao Wu 1004542140 #Yaozhong Zhang 1003915409

Abstract

As the society grows, people define happiness in more various way than before. While in the past, peace and stable life are probably the two criteria that define people's feeling of life, the number of such criteria increases since the society are growing in a relatively peaceful and technological way. The paper is focusing on the group of family who are included in the 2017 General Social Survey (GSS), collecting basic data and investigating the potential relationship between feeling of life and age, number of children per family, family annual income. Professor Samantha-Jo Caetano and Rohan Alexander provides a thorough method for cleaning the raw data by cleaning the variables' name and change values into numbers. Our null hypothesis is that age, number of children per family, family annual income does not have correlation with feeling of life. Based on the clean data, the linear regression model is used to demonstrate the correlation between feeling of life and age, number of children per family, family annual income. At the same time, the significance of each variables are also tested by function summary. The result shows two aspects. Age and the number of children are significant enough to reject the null hypothesis even though they shows a slightly positive correlation with feeling of life while feeling of life is negatively related with the different family annual income groups. The result also shows that family with higher income are less negatively affected by the increase of the children number. Therefore, the paper could draw the conclusion that age and total children in the family sightly and positively affects feeling of life while family income are showing much greater negative impact on the feeling of life. Last but not least, the more a family earned, the less the family will be negatively affected by the number of children in the family.

Introduction

Feelings of life always depends heavily on one's age, family and income. Usually people in different ages would have different feels about their life. Moreover, individuals who have more kids, or higher income, would differ a lot in their feelings for life. Therefore, this data analysis is about finding the potential relationship between respondents' self feelings of life and their age, total children have in family and family income. An multiple regression will be implemented to find the relation. The remainder of the paper is organized as follows. In Section 2 (Data section), we will introduce the original data and why we choose the data. In Section 3 (Model section), we will develop a model to explain the relation between feelings for life and age, family, income. In Section 4 (Results section), model results will be shown, as well as interpretations for the results. Section 5 (Discussion, weakness section) concludes and discuss the weakness of the model. And the final section will be what can we do next.

Data

From the data collection GSS, the raw data of this project is based on the family survey of 2017. The raw contains 460 variables and over 20000 observations in total. This enormous data allows a great amounts of investments in the relationship between different variables. This means that the investment in those variables' relationships which remains unknown, is easier to carry out than the dataset with few variables. Therefore, more novel and interesting relationships between different variables can be discovered. Its large sample is one of the features that makes this family survey significant. Furthermore, one important feature of this dataset is that most of the values are numeric, which makes the data cleaning and coding easier. However, the raw data has variables with weird names, which could be the drawback to the data. Moreover, existence of missing values are also the potential drawbacks. Fortunately, after applying gss_cleaning.R code, the clean_data.csv contains only 81 variables with them nicely named. This ensured that the building of the linear regression model is clear to create and reproduce.

Model

We are using a linear regression model method in R to find whether there is a linear relationship on a respondents' age, total children have in family and their family income affects their self feelings of life. In the beginning, we assume our null hypothesis that there is no linear relationship between age, total children, family income to their respondent's feelings of life, as beta = 0. and our alternative hypothesis is that there is a linear relationship, as beta not equal to 0.Since all the explanatory variables(age,total children and income) are not random variables, and the output variable in model (feelings of life) is in a range of 1-10 in survey, it leads me to choose the simple linear regression model than other types of models as figures:

```
##
## Call:
## lm(formula = feelings_life ~ age + total_children + as.factor(income_family),
##     data = adata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5679 -0.7279  0.0488  1.2168  2.8423
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             7.9866316  0.0470850 169.622
## age                                     0.0055874  0.0007317   7.637
## total_children                          0.0611704  0.0085335   7.168
## as.factor(income_family)$125,000 and more  0.0675315  0.0420429   1.606
## as.factor(income_family)$25,000 to $49,999 -0.5086741  0.0430561 -11.814
## as.factor(income_family)$50,000 to $74,999 -0.2773256  0.0439947  -6.304
## as.factor(income_family)$75,000 to $99,999 -0.1607216  0.0459204  -3.500
## as.factor(income_family)Less than $25,000  -0.9127204  0.0467896 -19.507
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## age                                     2.33e-14 ***
## total_children                          7.85e-13 ***
## as.factor(income_family)$125,000 and more  0.108235
## as.factor(income_family)$25,000 to $49,999  < 2e-16 ***
## as.factor(income_family)$50,000 to $74,999 2.97e-10 ***
## as.factor(income_family)$75,000 to $99,999 0.000466 ***
## as.factor(income_family)Less than $25,000   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.609 on 20304 degrees of freedom
##   (290 observations deleted due to missingness)
## Multiple R-squared:  0.04361,    Adjusted R-squared:  0.04328
## F-statistic: 132.3 on 7 and 20304 DF,  p-value: < 2.2e-16
```

Results

The results of my linear regression model shows that age is positively related with a respondent's feelings of life, as for every 1 year older the respondent gets, the scale of feelings of life increases by approximately 0.0056 in scale of 1 to 10 while keep other variables fixed. Meanwhile, there is also a positive linear relationship between the total children a respondent has and her feelings of life. By additional 1 children in a respondent's family, the feeling of life scale increases by approximately 0.061 keeping other variables fixed. For the relationship between a respondent's total family income and her feeling of life scale, the different income groups varies in results while keep other variables fixed. With the respondent's family total income is 125,000 dollars and higher, since the p-value of this estimate is 0.108235, greater than 0.1, we have no evidence rejects the null hypothesis that there is no linear relationship between family income and feelings of life.

For respondents' family income in between 25,000 dollars and $49,999, there is a negative relationship between their family income and their feelings of life, as if they have total family income in this range, their feelings of life scale decreases by a factor of 0.509. If a respondent's total family income is between 50,000 dollars and 74,999 dollars, their feelings of life scales decrease by a factor of 0.277. Relatively, if a respondent's total family income is between 75,000 dollars and 99,999 dollars, their feelings of life scale decreases by a factor of 0.161, and if a respondent's total family income is less than 25,000 dollars, their feeling of life scale decreases by 0.913. Note that all the factor of scale is rounded with 3 decimal places.

The data has a residual error of 1.609 with 20304 degrees of freedom and there are 290 un-responses. The coefficient of determination is 0.04361 which also illustrates that there is a posistive linear relationship.

Discussion

The result shows how age, total children and family annual income correlate with feeling of life. According to the summary, in terms of age, it shows a weak correlation with feeling of life. That means as the age increases, individuals do not significantly feel an increase in feeling of

life. Similar to age, the number of children in each family shows low positive correlation with feeling of life even though the estimate is 10 times more significant than the age variable. However, both p values of the two variables shows that we have strong evidence to reject the null hypothesis that age and number of children have no effect on feeling of life even though the effect is subtle.

More importantly, however, we see great correlation between different family income groups and feeling of life. Except for the group earning 125,000 dollars and more whose have no evidence to show this stage of income have any relation with feeling of life, other groups show not only relatively significant correlations with feeling of life, but also display a pattern of the correlations. Among the group earning less than 125,000 dollars, more children lead to less feeling of life. More interestingly, as families earned less annually, the stronger negative correlation is shown with feeling of life, which means in a poorer family, having more children may lead to more significant decrease in feeling of life. This is reasonable that with less income, it is harder to keep all children fed and educated well. In this case, income is the threshold for increasing family's feeling of life. Compared with families with higher income, their feeling of life is restricted by money. For families with higher income, there may exist more variables that affect their feeling of life rather than the household cost.

Weaknesses

The weakness of this data analysis would be the following. First, the data is incomplete. As mentions above, there are many NA in the data set, which means people did not fully answer the survey questions. This may result inaccuracy while working on data analysis with a clean data set. Secondary, since this is a survey, the information that provided on the data set might not be accurate. While looking at the data, most of the variables are related to their personal information, so that the data collected from this data set is really based on their personal feeling. For example, variables like self_rated_health and self_rated_mental_health could not give us a accurate measure of the realistic health and mental health of each observation since it' depend on their personal feeling about themselves when answering these two variables.

Moreover, the information that provided on income_family may not be that accurate because people may not have time to calculate their actual income when doing the survey. The inaccuracy of information of this data set would also be a weakness for this data analysis. Moreover, here we are looking for how would the feeling of life be affected by the age, total number of children and income in each family. However, the estimate value of age and total children are very small, which can't dertermine that whether there are relationship amoung age, total children and feeling of life. That is a weakness of our study.

Next Steps

Next steps for our research would be building up a survey related to all three factors: the respondent's age, family income and total children with respect to how the respondent's feelings of life, but with smaller target population rather than the census methods. For instance, we are aiming to use a total groups of people from my neighborhoud who is a Canadian citizens as a sample. The sampling method could also be applied here as an potential improvement for our algorithm. Our further research may also consider to find a different model approach that fits the new survey data rather than simple linear regression models. A bayesian models approach may also apply here for our futher research on estimating the Canadian people's feelings of life.

References

1. General Social Survey, Cycle 31: 2017: Family. Chass Data centre. Retrieved from: http://www.chass.utoronto.ca/.
2. Rohan Alexander, Samantha-Jo Caetano. gss_cleaning.R.