Statistical Inference on How Age, Total Number of Children Owned and Family Income Affect a Person's Feelings about Life as A Whole

Group 3

2020/10/19

Statistical Inference on How Age, Total Number of Children Owned and Family Income Affect a Person's Feelings about Life as A Whole

Xiaoxuan Han, Yicheng Ding, Yimeng Ma, Sun Hao

2020/10/19

Code and data supporting this analysis is available at: https://github.com/YichengDing0106/PS2-Group3.git

Abstract

The data of our group report is about the living conditions and well-being of Canadians gathered by the General Social Survey program. Based on the data sets, we wondered how the family income, age and total children owned affect the Canadians' feelings (good or bad) which is a binary response variable. Thus, we established a multiple logistic regression model to make a statistical analysis on this. Finally, since we found that all p-values in our models were less than 0.05, we reject the null hypothesis that there is no relationship between dependent variables(family income, age and number of total kids) and independent variables(respondents' feelings), and we are able to predict a person's feeling by using age, total number of children owned and family income after fitting the logistic regression model.

Introduction

With the development of society, more and more living conditions, such as family income, religion, average worked hours per day and total number of children, affect the well-being of people. Well-being of citizens is able to influence many aspects of this country such as economic development, social development and ability of leaders. The data provided by General Social Survey (GSS) program collected multiple different living conditions of Canadians and their feelings about life, which provide an opportunity for us to statistically analyze the factors affecting well being of Canadians. In our report, the purpose is investigating the effect of three variables (age, family income and total children owned) to the respondents' feelings. Thus, the null hypothesis of our report is that there is no relationship between age, family income and total children owned and their feelings. If the results show significant relationship between our dependent variable and independent variables, the country and government or even our citizens would have a more accurate and clear directions and consider more targeted methods about how to improve the well beings of Canadians. For the analysis, We established two models, simple logistic regression model and multiple logistic regression model to

statistically present the relationship between these variables. Our result shows that all p-values of coefficients are less than 0.05, in other worlds, that is we can reject the null hypothesis, and conclude that there is a significant relationship between age, family income and total children owned and respondent's feelings.

Data

The data set we used is the General Social Survey about changes in Canadian families in 2017. It covers information about marriage, parents' history, family origins, children's situation, income, and other social characteristics. The target population is all non-institutionalized persons 15 years of age or older and live within Canada. The frame this survey used is the combined landline and cellular telephone numbers from the Census and some sources with Statistics Canada's dwelling frame. In the frame, there are groups of several telephone numbers associated with the same address. The sample was representative of all households in the 10 provinces. All information is collected from a random household member 15 years of age or older, and proxy responses are not allowed. If there is a non-response, a series of adjustments to the survey weights will be implement to account for non-response as much as possible.

The sample is based on a stratified random sampling. 15 CMAs (Census Metropolitan Area) such as Montreal, Quebec City, Toronto, Ottawa are each considered separate strata. Three more strata were formed by grouping the remaining CMAs in each of Quebec, Ontario and British Columbia. The non-CMA areas of each provinces were grouped to be 10 more strata. Hence, there are 27 strata in total. There are determined minimum sample sizes for each province. Once the sample size targets had been met, the remaining sample was allocated to the strata in a balance way.

Data was collected from survey respondents using telephone interviewing, so it may has some non-response cases. Questions about income shows rather high non-response rates and the incomes respondents reported are usually rough estimate which lack accuracy. There exists sample error because the data are based on a sample of persons. The overall response rate is 52.4% which is not too high. Moreover, households without telephones were not covered. These non-response and non-covered cases may cause the non-sampling error.

The variables we choose are feelings about life, age, numbers of children, and family income. Since "feelings about life as a whole" have 11 levels group, but by observing the data set, we notice that there is not too much difference between adjacent groups, so we decide to divide these 11 levels into two groups: one group of respondents who have feelings about life level below the mean of feelings level overall, and one group of respondents who have feelings about life level above mean of feelings level. We choose mean as the cutoff point since it can better represent the average feelings about life level among all the respondents. Thus, new variable is created as "good_or_bad". Since "good_or_bad" has two levels, we encode this as one dummy variable.

```
## # A tibble: 20,312 x 6
##    caseid feelings_about_li~   age family_income      total_children good_or_bad
##     <int>              <int> <dbl> <chr>                       <int> <chr>
##  1      1                  8  52.7 $25,000 to $49,999              1 1
##  2      2                 10  51.1 $75,000 to $99,999              5 0
##  3      3                  8  63.6 $75,000 to $99,999              5 1
##  4      4                 10  80   $100,000 to $ 124~              1 0
##  5      5                  8  28   $50,000 to $74,999              0 1
##  6      6                  9  63   $50,000 to $74,999              2 0
##  7      7                  4  58.8 Less than $25,000               2 1
##  8      8                 10  80   Less than $25,000               7 0
##  9      9                  8  63.8 Less than $25,000               0 1
## 10     10                  5  25.2 Less than $25,000               1 1
## # ... with 20,302 more rows
```

```
## [1] 20312     6
```

There are 20602 observations in total in the origin data set. After removing all NAs, the data we will use for this analysis has 20312 observations in total.

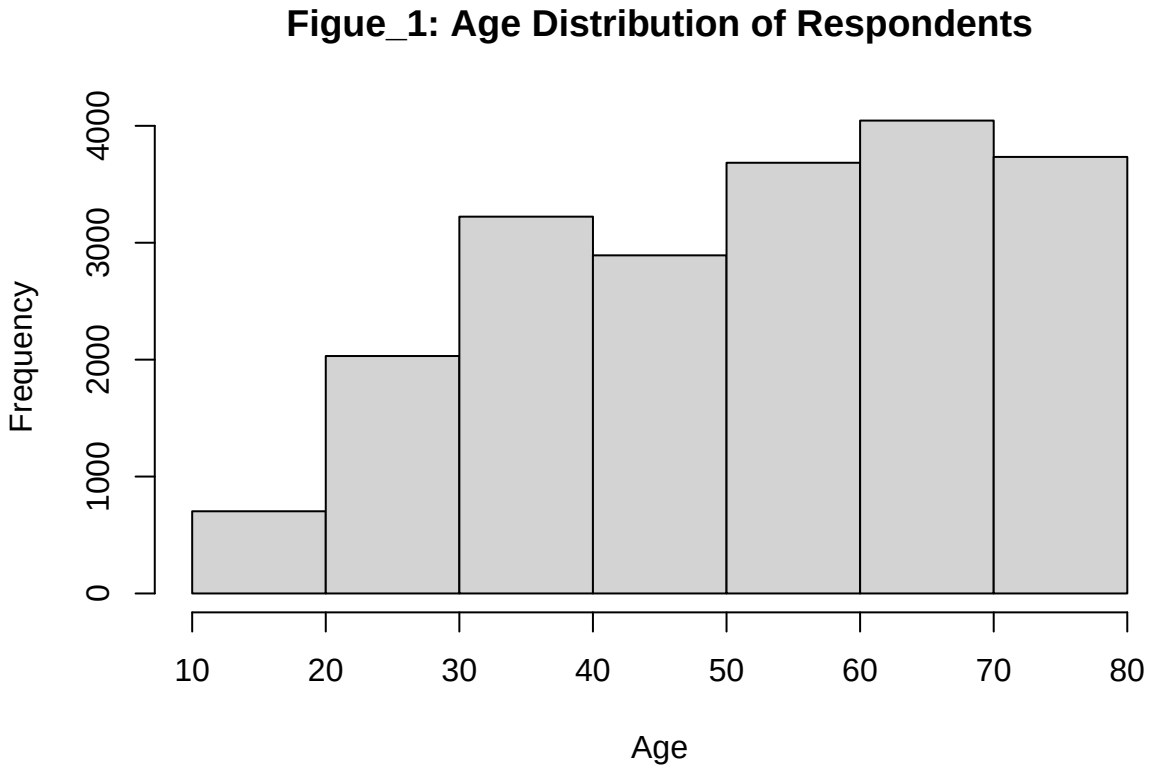


Figure 1 is age distribution of all respondents. From the figure, we can find that it is a left skewed distribution. Most respondents are 30 years of age or older and just a few respondents from age 10 to 30.

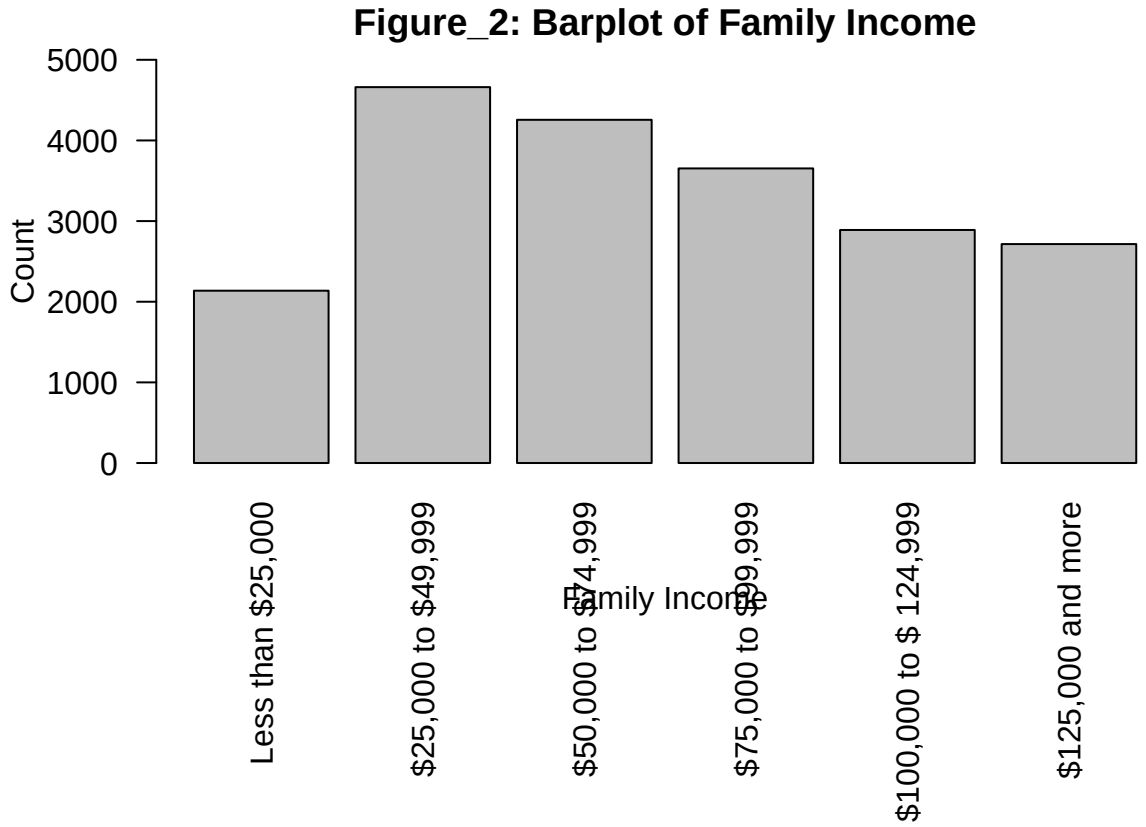


Figure 2 is the bar plot of family income which is a right skewed distribution. Most respondents have family income between $25000 to $100000 and few respondents have family income less than $25000.

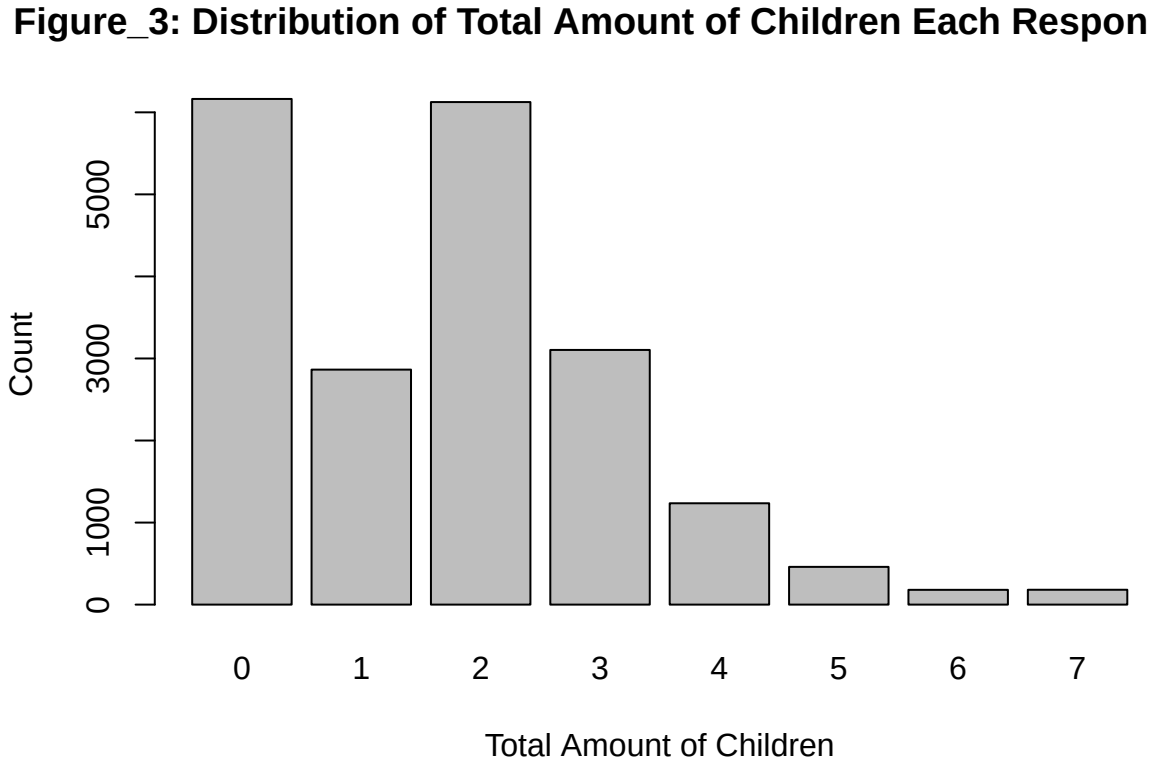


Figure 3 is the distribution of total amount of children owned by each repsondent. We can find that it is a

right skewed distribution. Most respondents have 0 to 3 children and few respondents have 6 to 7 children.

Figure_4: Barplot of Feeling about Life as a Whole

Figure_5: Barplot of Feeling about Life (Good or Bad)

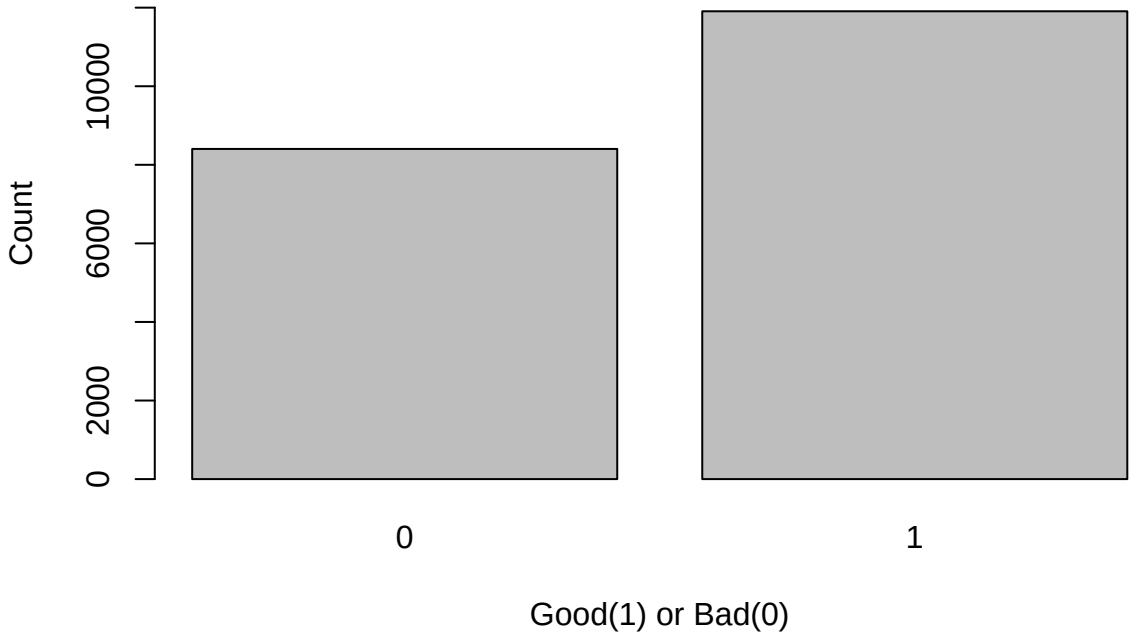


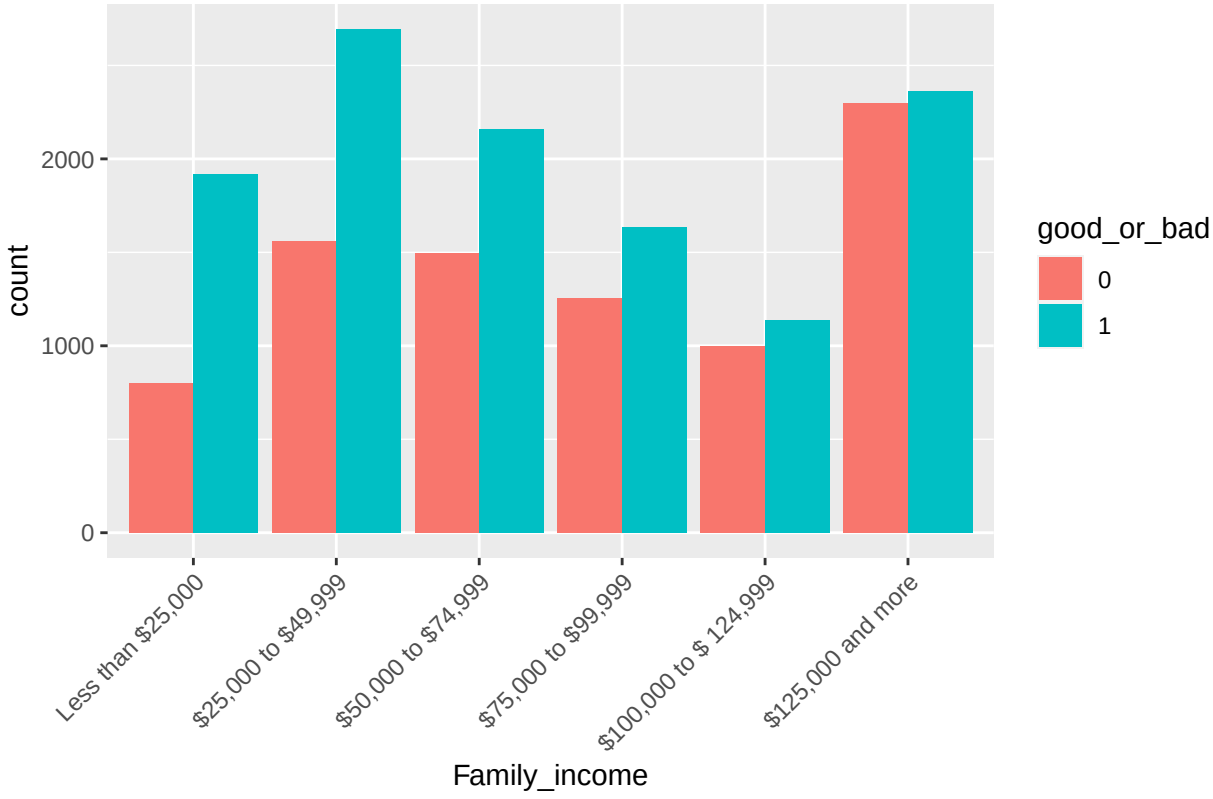
Figure 4 is a bar plot of feeling about life as a whole, which can be translated into "Are you satisfied with your life?". From level 0 - 10, satisfaction increased gradually, so 0 means very dissatisfied and 10 means very satisfied. The distribution for this plot is left skewed which means most respondents are satisfied with their

lives. Only a few respondents choose under level 5. Most respondents chose 8 which means they think their lives can be further improved.

In figure 5, these respondents are divided into two group - Good(1) and Bad(0). If a respondent's answer is lower than the mean of feeling about life, he will be put into group of Bad(0) which means his satisfaction with life is lower than average and vice versa. After doing this, we can find that most respondents' satisfaction with life are higher than average.

```
##                     Var1 Var2 Freq
## 1  $100,000 to $ 124,999    0  998
## 2      $125,000 and more    0 2297
## 3     $25,000 to $49,999    0 1562
## 4     $50,000 to $74,999    0 1494
## 5     $75,000 to $99,999    0 1254
## 6      Less than $25,000    0  799
## 7  $100,000 to $ 124,999    1 1139
## 8      $125,000 and more    1 2364
## 9     $25,000 to $49,999    1 2694
## 10    $50,000 to $74,999    1 2159
## 11    $75,000 to $99,999    1 1636
## 12     Less than $25,000    1 1916
```

Figure_6: Barplot of Renspondent's Feeling and Family Income



In figure 6, we combined family income and feeling about life and get a new bar plot. We can find that for respondents with more family income, good group and bad group almost have the same amount of respondents. However, for respondents with less family income, there are more respondents in good group than bad group. As the higher the income, the smaller the gap between good group and bad group.

Model

In statistics, logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Hence, The model we used is logistic regression model throughout our analysis.

First, we wish to see explore our data and see if only family income itself would affect the respondent's feeling (good or bad). Since respondent's feeling is a binary response categorical variable and family income is gathered as different groups, which is an independent variable, a simple logistic regression model is appropriate in this case.

```
##
## Call:
## glm(formula = as.numeric(good_or_bad) ~ as.factor(family_income),
##     family = "binomial", data = feelings_life_data.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5641  -1.2922   0.8349   1.0668   1.1652
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               0.13215    0.04336   3.048   0.0023
## as.factor(family_income)$125,000 and more -0.10340    0.05233  -1.976   0.0482
## as.factor(family_income)$25,000 to $49,999  0.41291    0.05377   7.679  1.6e-14
## as.factor(family_income)$50,000 to $74,999  0.23604    0.05489   4.300  1.7e-05
## as.factor(family_income)$75,000 to $99,999  0.13376    0.05735   2.333   0.0197
## as.factor(family_income)Less than $25,000   0.74248    0.06044  12.284  < 2e-16
##
## (Intercept)                              **
## as.factor(family_income)$125,000 and more  *
## as.factor(family_income)$25,000 to $49,999 ***
## as.factor(family_income)$50,000 to $74,999 ***
## as.factor(family_income)$75,000 to $99,999 *
## as.factor(family_income)Less than $25,000  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27551  on 20311  degrees of freedom
## Residual deviance: 27198  on 20306  degrees of freedom
## AIC: 27210
##
## Number of Fisher Scoring iterations: 4
```

The mathematical model equation obtained is

$$log(p/(1-p)) = 0.13215 - 0.1034 * X_1 + 0.41291 * X_2 + 0.23604 * X_3 + 0.13376 * X_4 + 0.74248 * X_5$$

In this model, p is the probability that a person feels good. X_1 represents family income group 125,000 and more dollars, X_2 represents family income group 25,000 to 49,999 dollars, X_3 represents family income group 50,000 to 74,999 dollars, X_4 represents family income group 75,000 to 99,999 dollars, and X_5 represents family income group Less than 25,000 dollars.

We will interpret the if there is a group having income $125,000 and more, the log odds of feelings (good or bad) will change by -0.10340 (which means it is decreased by 0.10340) when other variables hold constant.

Similar interpretation will applied to other family income groups as well.

Since the p values of coefficient are all smaller than siginificance level 0.05, we can conclude that there is a relationship between the respondent's family income and the respondent's feeling.

However, it is very unlikely that a respondent's feeling is only associated with the respondent's family income. Also, because a model with more variables presents less statistical power, we eventually selected two more variables, age and total children owned, which are also likely to affect the respondent's feeling. We will then fit a logistic regression model of multiple explanatory variables (including both numerical and categorical), which is called multiple logistic regression, since the respondent's feeling (good or bad) is still a binary response variable as the dependent variable, but we have three explanatory variables in the model now. This is the model we will focus on in this statistical analysis. The software that will be able to run this model is Rstudio.

In this multiple logistic regression model, we use age as the numerical explanatory variable rather than using age group, which is a categorical variable. This is because when predicting the probability of feelings about life that is above average (good), numerical variable will make the value of the result become more precise, and we can then get a better predicted value. Also, if we use age group, say each group has 10 years of range, then the result generated by this will likely to be not convincing enough since people who are 40 years old probably have different feelings about life than people who are 50 years old. Total number of children is a numerical explanatory variable as well, since it consists of continuous integers. We use the exact number of total number of children instead of splitting them into few groups because by observing the data, people normally have 0 to 7 children. It is meaningless to split them into groups since having one more child might have a large effect on a person's life, so the difference between each successive number of children owned matters a lot. Lastly, we use family income as a categorical variable, where incomes are split into groups of different ranges. We use income groups because that is what been given to us in the General Social Survey-Family 2017 data set.

```
## [1] 0.4470329
```

```
##
## Call:
## glm(formula = as.factor(family_income) ~ as.numeric(total_children),
##     family = "binomial", data = feelings_life_data.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1244   0.4703   0.4710   0.4718   0.4753
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.146001   0.034467  62.263   <2e-16 ***
## as.numeric(total_children) -0.003194   0.015349  -0.208    0.835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13665  on 20311  degrees of freedom
## Residual deviance: 13665  on 20310  degrees of freedom
## AIC: 13669
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = as.factor(family_income) ~ as.numeric(age), family = "binomial",
##     data = feelings_life_data.clean)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2707   0.4169   0.4523   0.5028   0.5774
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.524131   0.067025  22.740   <2e-16 ***
## as.numeric(age) 0.012187   0.001284   9.494   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13665  on 20311  degrees of freedom
## Residual deviance: 13575  on 20310  degrees of freedom
## AIC: 13579
##
## Number of Fisher Scoring iterations: 5
```

In this multiple logistic regression model, convergence will occur quickly because our model is a simple model, and the variables we choose to estimate the model are not highly correlated. We check this by computing correlation and finding out association between each variable. It turns out that the correlation between age and children is about 0.45, which is not high, and there turns out to be no association between family income and total children owned since the p value is 0.835 when conducting a hypothesis test on their simple linear regression model, so they are not highly correlated either.

```
##
## Call:
## glm(formula = as.factor(good_or_bad) ~ age + as.factor(family_income) +
##     as.numeric(total_children), family = "binomial", data = feelings_life_data.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7651  -1.2493   0.8626   1.0445   1.4027
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                 0.6430917  0.0601176  10.697
## age                                        -0.0080656  0.0009466  -8.521
## as.factor(family_income)$125,000 and more  -0.1165019  0.0525808  -2.216
## as.factor(family_income)$25,000 to $49,999  0.4887879  0.0546476   8.944
## as.factor(family_income)$50,000 to $74,999  0.2800365  0.0554078   5.054
## as.factor(family_income)$75,000 to $99,999  0.1495707  0.0576422   2.595
## as.factor(family_income)Less than $25,000   0.7990926  0.0612128  13.054
## as.numeric(total_children)                 -0.0698405  0.0108475  -6.438
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## age                                         < 2e-16 ***
## as.factor(family_income)$125,000 and more   0.02671 *
## as.factor(family_income)$25,000 to $49,999  < 2e-16 ***
## as.factor(family_income)$50,000 to $74,999 4.32e-07 ***
## as.factor(family_income)$75,000 to $99,999  0.00946 **
## as.factor(family_income)Less than $25,000   < 2e-16 ***
## as.numeric(total_children)                 1.21e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27551  on 20311  degrees of freedom
## Residual deviance: 26991  on 20304  degrees of freedom
## AIC: 27007
##
## Number of Fisher Scoring iterations: 4
```

```
##                                             Estimate   Std. Error   z value
## (Intercept)                                 0.643091706 0.0601176193 10.697225
## age                                        -0.008065642 0.0009465892 -8.520742
## as.factor(family_income)$125,000 and more  -0.116501929 0.0525807868 -2.215675
## as.factor(family_income)$25,000 to $49,999  0.488787886 0.0546476001  8.944361
## as.factor(family_income)$50,000 to $74,999  0.280036458 0.0554077765  5.054100
## as.factor(family_income)$75,000 to $99,999  0.149570735 0.0576422346  2.594812
## as.factor(family_income)Less than $25,000   0.799092588 0.0612128479 13.054328
## as.numeric(total_children)                 -0.069840539 0.0108474906 -6.438405
##                                                Pr(>|z|)
## (Intercept)                                1.048718e-26
## age                                        1.585345e-17
## as.factor(family_income)$125,000 and more  2.671378e-02
## as.factor(family_income)$25,000 to $49,999 3.741065e-19
## as.factor(family_income)$50,000 to $74,999 4.324244e-07
## as.factor(family_income)$75,000 to $99,999 9.464280e-03
## as.factor(family_income)Less than $25,000  6.003753e-39
## as.numeric(total_children)                 1.207354e-10
```

The mathematical model equation obtained from table_3 for this multiple logistic regressionn is

$$log(p/(1-p)) = 0.6430917 - 0.0080656 * Age - 0.1165019 * X_1 + 0.4887879 * X_2 + 0.2800365 * X_3 + 0.1495707 * X_4 + 0.7990926 * X_5 - 0.$$

In this model, p is the probability that a person feels good. X_1 represents family income group 125,000 and more dollars, X_2 represents family income group 25,000 to 49,999 dollars, X_3 represents family income group 50,000 to 74,999 dollars, X_4 represents family income group 75,000 to 99,999 dollars, and X_5 represents family income group Less than 25,000 dollars.

From the above table (table_3), it can be explained that for every additional unit increase in age, we expect the log odds of the respondent's feeling to decrease by 0.0080656 when all the other variables hold constant; for every additional unit increase in the amount of total children the respondent has, we expect the log odds of the respondent's feeling to decrease by 0.0698405 when all the other variables hold constant.

The categorical variables for family income have a slightly different interpretation. For example, we will interpret the if there is a group having income $125,000 and more, the log odds of feelings (good or bad) by -0.1165019 (which means it is decreased by 0.1165019) when other variables hold constant. Similar interpretation will applied to other family income groups as well.

We will conduct a hypothesis test at 0.05 significance level on finding if there is any significant relationship between the dependent variable, respondent's feeling (good or bad), and the independent variables, age, family income, and total children owned. The null hypothesis is that there is no relationship, meaning all the coefficients β_i, where i is from 1 to 7, are zero. The alternative hypothesis is that there is a relationship, meaning at least one of the coefficients β_i, where i is from 1 to 7, is not equal to zero. Then, we will also generate the confidence intervals for each coefficient to further support the conclusion of our hypothesis test.

```
## Waiting for profiling to be done...
```

```
##                                                    2.5 %       97.5 %
## (Intercept)                                     0.525401905  0.761075096
## age                                            -0.009922351 -0.006211648
## as.factor(family_income)$125,000 and more      -0.219620341 -0.013491425
## as.factor(family_income)$25,000 to $49,999      0.381697975  0.595928758
## as.factor(family_income)$50,000 to $74,999      0.171440301  0.388650629
## as.factor(family_income)$75,000 to $99,999      0.036594641  0.262564178
## as.factor(family_income)Less than $25,000       0.679304606  0.919274800
## as.numeric(total_children)                     -0.091104957 -0.048580672
```

Last, to better see how our model fits, we will perform another test focusing on whether the model with predictors fits significantly better than a model with just an intercept (i.e., a null model). This test is based out the output produced by summary(mylogit), which includes the coefficient and all null and deviance residuals and AIC. The test statistic is the difference between the residual deviance for the model with predictors and the null model. The test statistic is distributed chi-squared with degrees of freedom equal to the differences in degrees of freedom between the current and the null model (i.e., the number of predictor variables in the model).

```
## [1] 559.426
```

```
## [1] 7
```

```
## [1] 1.322119e-116
```

There are some advantages and disadvantage of fitting a logistic regression model on finding out how age, total number of children owned and family income affect a person's feelings about life as a whole. As for the advantages, first, it has good accuracy for this simple dataset and it performs well since this dataset of General Social Survey is linearly separable. Second, it can interpret model coefficients as indicators of feature importance, where features include family income, total children owned and age in our case. Third, a logistic regression model is less inclined to overfitting when the data set is not high dimensional, so in our case, it is unlikely to overfit since the dimension is only three and it is low. As for the disadvantages, first, we will not get complex relationship among these variables when using logistic regression model since its algorithm is not powerful enough. Second, a logistic regression model automatically assumes linearity between the dependent variable and independent variables, whereas sometimes relationship of linearity is not true.

An alternative model for finding out how age, total number of children owned and family income affect a person's feelings about life as a whole is ordinal logistic regression model. An ordinal regression is a type of regression analysis used for predicting an ordinal variable, i.e. a variable whose value exists on an arbitrary scale where only the relative ordering between different values is significant. In our case, the variable feelings about life will not be split into two large groups, but remain as orders from 0 to 10. All the other variables will not change and the procedure will most likely assemble the normal multiple logistic regression model. One strength about this is that we can explore more deeply and make better predictions of each level of feelings about life. One drawback is that it some of the results of each feeling level might not be practical, since knowing the exact feelings level is not as meaningful as what the respondent feels about life comparing to the overall population.

Results

```
##
## Call:
## glm(formula = as.factor(good_or_bad) ~ age + as.factor(family_income) +
##     as.numeric(total_children), family = "binomial", data = feelings_life_data.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.7651  -1.2493   0.8626   1.0445   1.4027
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                                0.6430917  0.0601176  10.697
## age                                       -0.0080656  0.0009466  -8.521
## as.factor(family_income)$125,000 and more -0.1165019  0.0525808  -2.216
## as.factor(family_income)$25,000 to $49,999 0.4887879  0.0546476   8.944
## as.factor(family_income)$50,000 to $74,999 0.2800365  0.0554078   5.054
## as.factor(family_income)$75,000 to $99,999 0.1495707  0.0576422   2.595
## as.factor(family_income)Less than $25,000  0.7990926  0.0612128  13.054
## as.numeric(total_children)                -0.0698405  0.0108475  -6.438
##                                           Pr(>|z|)
## (Intercept)                                < 2e-16 ***
## age                                        < 2e-16 ***
## as.factor(family_income)$125,000 and more   0.02671 *
## as.factor(family_income)$25,000 to $49,999  < 2e-16 ***
## as.factor(family_income)$50,000 to $74,999 4.32e-07 ***
## as.factor(family_income)$75,000 to $99,999  0.00946 **
## as.factor(family_income)Less than $25,000   < 2e-16 ***
## as.numeric(total_children)                 1.21e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27551  on 20311  degrees of freedom
## Residual deviance: 26991  on 20304  degrees of freedom
## AIC: 27007
##
## Number of Fisher Scoring iterations: 4
```

From this table, it can be observed that all the p values are smaller than significance level 0.05, so we reject the null hypothesis that there is no relationship between the dependent variable, respondent's feeling, and the independent variables, age, family income, and total children owned, and we are able to support the alternative hypothesis.

```
## Waiting for profiling to be done...
```

```
##                                                2.5 %       97.5 %
## (Intercept)                                0.525401905  0.761075096
## age                                       -0.009922351 -0.006211648
## as.factor(family_income)$125,000 and more -0.219620341 -0.013491425
## as.factor(family_income)$25,000 to $49,999 0.381697975  0.595928758
## as.factor(family_income)$50,000 to $74,999 0.171440301  0.388650629
## as.factor(family_income)$75,000 to $99,999 0.036594641  0.262564178
## as.factor(family_income)Less than $25,000  0.679304606  0.919274800
## as.numeric(total_children)                -0.091104957 -0.048580672
```

From the confidence interval generated for each coefficient, 0 is not contained in any of the 95% confidence interval, so all the confidence intervals do not contain the null hypothesis value, which makes our results statistically significant.

```
## [1] 559.426
```

```
## [1] 7
```

```
## [1] 1.322119e-116
```

Based on the the result of the test focusing on whether the model with predictors fits significantly better than a model with just an intercept (i.e., a null model), we get a chi-square of 559.426 with 7 degrees of freedom and a p-value of 1.322119e-116 less than 0.05. We then conclude that our model as a whole fits significantly better than an empty model.

Discussion

Now, we come up with results conclusion and discussion part. First of all, we will demonstrate what we found from **Data** and **Model** section, and provides some comments on how it relates to the original goal of the study. Next, we will demonstrate some potential issues (*weaknesses of study*) that we found during the study. Finally, in order to make the study results more impressive and accurate, we also propose some solutions (*next steps*) to each corresponding weakness.

From **Data Analytical Results**, we conclude that most of people in the data set have *life of feeling* score greater than 8, that means more than half of people have good feeling of life (It can also be told in Figure_5). *Family income* between $25,000 to $49,999 has the biggest proportion in the data, and more people have good feeling of life if their family income in the range of $25,000 to $49,999.

From **Model Results**, we conclude that *life of feeling* has a relationship with *Age*, *Family income* and *Number of total children*. Among these three variables, Age and family income in the range of $25,000 to $49,999 have the biggest impact to *life of feeling*.

Weaknesses

There are also weaknesses and room of improvements that we found during the study. For example:

- In this study, we reclassified *life of feeling* into 'good' and 'bad' according to the mean of *life of feeling* scores. However, this rule may cause some non-sampling errors because we changed original survey's responses and create another new variable (*good_or_bad*) as research target.

- When fitting logistic regression models, we only consider impact of few different predictors. However, there are 80 possible predictors in the data set. It will cause model prediction error.

- In the GSS data set, there exists some mixed effects among different variables, but they are not considered in the study.

- With respect to study improvements, we showed few plots and model summaries. We are able to add more data visualizations in order to show study results more clearly.

Next Steps

From above section, we listed out few weaknesses and improvement items of the study. At the same time, we also proposed corresponding solutions to each of them.

- In order to eliminate biases from variable transformation. Instead of reclassifying *life of feeling* into 'good' and 'bad', we can use score of *life of feeling* in the original data set. And instead, Poisson regression model can be used for analyzing score of *life of feeling*.

- Fit regression model between *life of feeling* with all other variables (predictors) and using step-wise methods to select most significant predictors. It can make final model more accurate and less biases.

- Use **Generalized Linear Mixed Model** for regression model part. It will consider *Mixed Effects* for different categories, which can also reduce biases of model.

References

1. Statistics Canada. (Feb 7, 2019). General Social Survey - Family for 2017 (Cycle 31). Retrieved from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function= getSurvey&Id=335816
2. Craggs, Jason G. "x-Axis Labels Overlap - Want to Rotate Labels 45º." RStudio Community, 29 Apr. 2020, community.rstudio.com/t/x-axis-labels-overlap-want-to-rotate-labels-45/63800/2.
3. Craggs, Jason G. "x-Axis Labels Overlap - Want to Rotate Labels 45º." RStudio Community, 29 Apr. 2020, community.rstudio.com/t/x-axis-labels-overlap-want-to-rotate-labels-45/63800/2.
4. "HOME." IDRE Stats, stats.idre.ucla.edu/r/dae/logit-regression/.
5. Rawat, Akanksha. "Ordinal Logistic Regression." Medium, Towards Data Science, 6 Jan. 2019, towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5.