Exploration if There is a Linear Between Self Rated Health, Mental Health, Age and Family Income

Xu Xinyi

2020/10/18

Exploration if There is a Linear Between Self Rated Health, Mental Health, Age and Family Income

Xu Xinyi

2020/10/18

Code and data supporting this analysis is available at: https://github.com/xuxiny17/A2.git

Abstract

The data was from Canadian General Social Survey data done in 2017, I chose to analyze the relation ship between family income and self rated mental health and mental health and I expect that there might be a linear relationship between these variables, I suppose that the better the self rated health and mental health the higher the family income might be thus I fitted a simple linear regression model to investigate my hypothesis. However, through analysis, I noticed that the there is a bias that could be brought by age, thus I took age into account. I found that the p-value for the multiple linear regression is over 0.05 and R^2 is very small, thus I dropped the variable self rated health and did multiple linear regression models regarding income, self rated mental health and age and income with the other two variables respectively. After analysis, though the most p-values are below 0.05, indicating a significance, the R^2 value are all below 0.3, which means that the two variables have a very weak relationship, thus in conclusion, my hypothesis the higher the family income the better the self rated mental health and health does not hold by the analysis of 2017 data.

Introduction

The data set collected a lot of information from residents in Canada, and from all these variables, the income, mental health and health drew my attention as I have heard constantly from media that people tend to feel bad or even commit suicide while experiencing financial hard times. But would mental health affect income? I was curious and wish to do some investigations.

Thus I would like to analyze the relationship between the family income, self rated mental health and health and through brief analysis, I noticed that the variable age was largely involved in the data, thus I took age into account as well. I would use linear regression model to fit as I expect that the younger the respondent, the higher the individual rate on his/her health and metal health, the more this family might earn.

The result could be important, as if the hypothesis that an less a family earns the more likely an individual is experiencing an illness holds, this may allow researchers who study health related issue to find their target faster and easier, as well as give government an idea on what proportion of the population they could offer more aid to.

Here is where you should give insight into the setting and introduce the goal of the analysis. Here you can introduce ideas and basic concepts regarding the study setting and the potential model. Again, this is the

introduction, so you should be explaining the importance of the work that is ahead and hopefully build some suspense for the reader. You can also highlight what will be included in the subsequent sections.

Data

Data was cleaned up following the instruction of the gss_cleaning.R file.

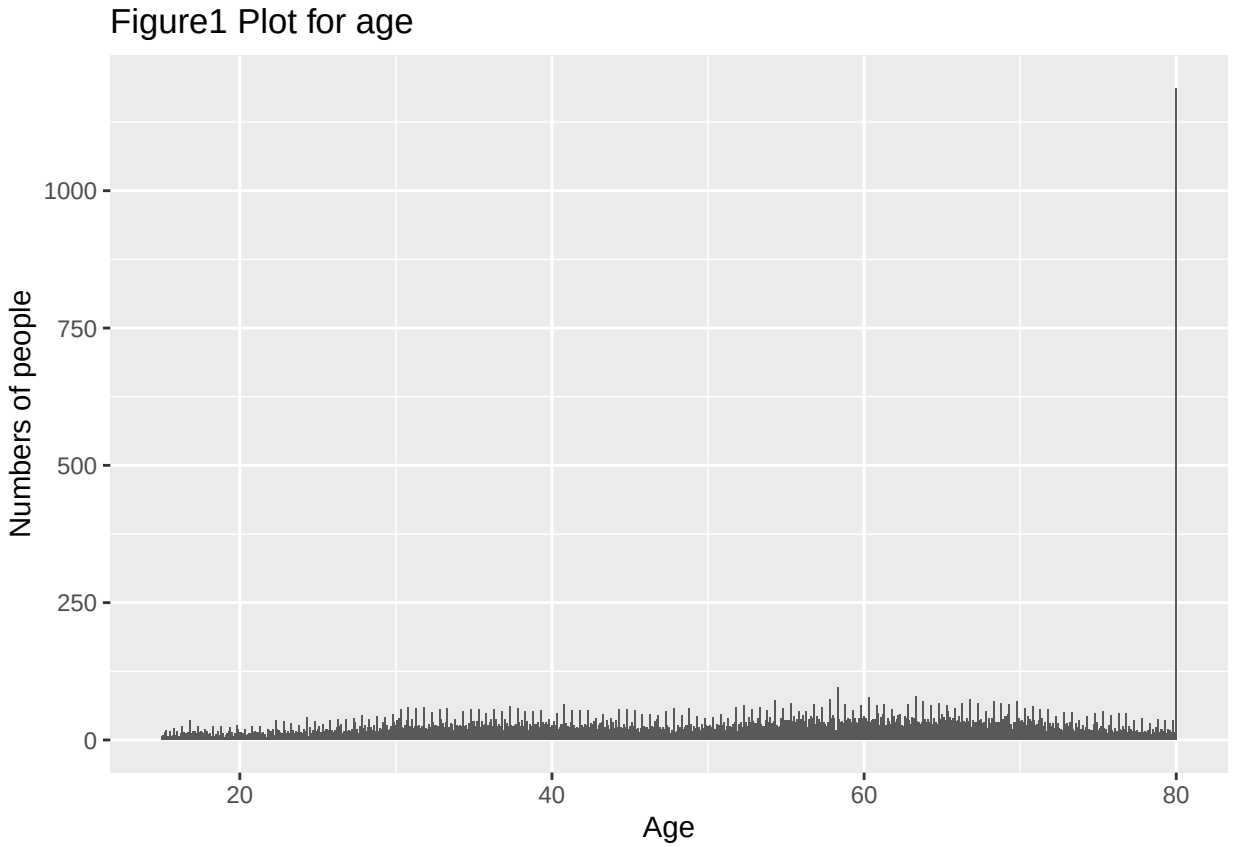
```
library(janitor)
library(tidyverse)
# install.packages("readr")
# install.packages("rvest")
library(rvest)
library(readr)

unprocessed <- read_csv("gss.csv")
glimpse(unprocessed)
```

```
## Rows: 20,602
## Columns: 81
## $ caseid                           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,...
## $ age                              <dbl> 52.7, 51.1, 63.6, 80.0, 28.0, 63.0...
## $ age_first_child                  <dbl> 27, 33, 40, 56, NA, 37, 40, 59, NA...
## $ age_youngest_child_under_6       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ total_children                   <dbl> 1, 5, 5, 1, 0, 2, 2, 7, 0, 1, 0, 0...
## $ age_start_relationship           <dbl> NA, NA, NA, NA, 25.3, NA, NA, NA, ...
## $ age_at_first_marriage            <dbl> NA, NA, NA, NA, NA, NA, NA, 22.1, ...
## $ age_at_first_birth               <dbl> 25.9, NA, 23.2, 27.3, NA, 25.8, 18...
## $ distance_between_houses          <dbl> 30, NA, NA, NA, NA, NA, NA, NA, NA...
## $ age_youngest_child_returned_work <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ feelings_life                    <dbl> 8, 10, 8, 10, 8, 9, 4, 10, 8, 5, 1...
## $ sex                              <chr> "Female", "Male", "Female", "Femal...
## $ place_birth_canada               <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_father               <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_mother               <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_macro_region         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ place_birth_province             <chr> "Quebec", "Ontario", "Ontario", "A...
## $ year_arrived_canada              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ province                         <chr> "Quebec", "Manitoba", "Ontario", "...
## $ region                           <chr> "Quebec", "Prairie region", "Ontar...
## $ pop_center                       <chr> "Larger urban population centres (...
## $ marital_status                   <chr> "Single, never married", "Married"...
## $ aboriginal                       <chr> "No", "No", "No", "No", "No", "No"...
## $ vis_minority                     <chr> "Not a visible minority", "Not a v...
## $ age_immigration                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ landed_immigrant                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ citizenship_status               <chr> "By birth", "By birth", "By birth"...
## $ education                        <chr> "High school diploma or a high sch...
## $ own_rent                         <chr> "Owned by you or a member of this ...
## $ living_arrangement               <chr> "Alone", "Spouse only", "Spouse on...
## $ hh_type                          <chr> "Low-rise apartment (less than 5 s...
## $ hh_size                          <dbl> 1, 2, 2, 2, 2, 2, 1, 1, 1, 6, 5, 1...
## $ partner_birth_country            <chr> "Canada", "Canada", "Canada", "Can...
## $ partner_birth_province           <chr> "Quebec", "Manitoba", "Ontario", "...
## $ partner_vis_minority             <chr> "Not a visible minority", "Not a v...
## $ partner_sex                      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
```

```
## $ partner_education                <chr> "Trade certificate or diploma", "B...
## $ average_hours_worked             <chr> "30.0 to 40.0 hours", "50.1 hours ...
## $ worked_last_week                 <chr> "Yes", "Yes", "No", "No", "No", "N...
## $ partner_main_activity            <chr> "Working at a paid job or business...
## $ self_rated_health                <chr> "Excellent", "Good", "Very good", ...
## $ self_rated_mental_health         <chr> "Excellent", "Good", "Good", "Very...
## $ religion_has_affiliation         <chr> "Has religious affiliation", "Don'...
## $ regilion_importance              <chr> "Somewhat important", "Don't know"...
## $ language_home                    <chr> "French", "English", "French", "En...
## $ language_knowledge               <chr> "French only", "English only", "Bo...
## $ income_family                    <chr> "$25,000 to $49,999", "$75,000 to ...
## $ income_respondent                <chr> "$25,000 to $49,999", "Less than $...
## $ occupation                       <chr> "Sales and service occupations", "...
## $ childcare_regular                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ childcare_type                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ childcare_monthly_cost           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ever_fathered_child              <chr> NA, "Yes", NA, NA, "No", NA, NA, N...
## $ ever_given_birth                 <chr> "Yes", NA, "Yes", "Yes", NA, "Yes"...
## $ number_of_current_union          <chr> NA, NA, NA, NA, "Second union", NA...
## $ lives_with_partner               <chr> "No", "No", "No", "No", "Yes", "No...
## $ children_in_household            <chr> "No child", "No child", "No child"...
## $ number_total_children_intention  <dbl> NA, NA, NA, NA, 2, NA, NA, NA, NA,...
## $ has_grandchildren                <chr> "No", "Yes", "Yes", "No", "No", "Y...
## $ grandparents_still_living        <chr> "No", "No", "No", "No", "Yes", "No...
## $ ever_married                     <chr> "No", "Yes", "Yes", "Yes", "No", "...
## $ current_marriage_is_first        <chr> NA, "Yes", "Yes", "Yes", NA, "Yes"...
## $ number_marriages                 <dbl> 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0...
## $ religion_participation           <chr> "Once or twice a year", "Don't kno...
## $ partner_location_residence       <chr> "In the same province", NA, NA, NA...
## $ full_part_time_work              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ time_off_work_birth              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ reason_no_time_off_birth         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ returned_same_job                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ satisfied_time_children          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ provide_or_receive_fin_supp      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_child_supp              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_child_exp               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_lump                    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_other                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_agreement               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ future_children_intention        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ is_male                          <dbl> 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0...
## $ main_activity                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ age_diff                         <chr> NA, "Respondent is 4 years older",...
## $ number_total_children_known      <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1...
```

Quick Age analysis for the data from GSS



The data I chose is the GSS data set from 2017. The data set is very interesting as it listed 20602 observations and gathered information on the individual's income, self rated health, living conditions, religion affinity and lots of other variables. It involves a lot of interesting factors and information that drives my attention. Some variables like religion affinity, self rated mental health, religion importance and income soon caught my eyes, there are lot that we could dive into and do investigation upon.

However, for many variables, even at the first glace we could witness that there are lots of data that is NA, which that the analyze done on those variables could be biased as some extreme values or important sets of data could be missing which could be a serious drawback. Except for that, the remaining values are listed out clearly and are all capable for usage.

Besides, from Figure1, we could witness that a large proportion of the data was collected from elderly people who is 80 years old, largely out weigh the people from other ages. This could also be a severe drawback as this would lead to serious bias, as the conclusion we got may not be accurate enough if we do not take the ages into account.

Methodology

The data was collected by GSS in 2017, and as mentioned in User's Guide (Format pdf) on the GSS website, the data was collected using a new frame which combines the telephone numbers with the Statistics Canada's Address Register, and eventually collects the data via telephone.

The **population** are people who live in Canada, while the **target population** is mentioned in the Guide: "all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada", the **frame population** is all the people who are non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada and answered the phone call, the **sample population** is all the who answered the call completed at least part of the survey.

The sampling approach they used is probably random sampling, but the trade off could be not every people who they tend to survey would answer the phone or answer all the question they ask. This may be the reason

why there a lot more elderly people than the younger ones, as they tend to have more time and patience to answer the phone. The non-response value are left NA in the data set and while we wish to do some analysis we should omit them. The cost could be time conducting the survey as calling each individual one by one is time consuming, and human labor who involved in conducting the survey.

Model

```
## tibble [20,496 x 2] (S3: tbl_df/tbl/data.frame)
##  $ income_family           : Factor w/ 6 levels "$100,000 to $ 124,999",..: 3 5 5 1 4 4 6 6 6 6 ...
##  $ self_rated_mental_health: Factor w/ 6 levels "Don't know","Excellent",..: 2 4 4 6 4 6 5 6 6 5 ...
##  - attr(*, "na.action")= 'omit' Named int [1:106] 461 791 1026 1443 1530 1634 1914 1921 2200 2485 ..
##   ..- attr(*, "names")= chr [1:106] "461" "791" "1026" "1443" ...
```

Income ~ Self rated mental health + Self Rated Health + Age

```
##
## Call:
## lm(formula = as.numeric(income_family) ~ as.numeric(self_rated_mental_health) +
##     as.numeric(self_rated_health) + as.numeric(age), data = data_reformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7157 -1.3287 -0.2863  1.4325  2.9448
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        2.893845   0.047835  60.497   <2e-16 ***
## as.numeric(self_rated_mental_health)  0.013133   0.007101   1.850   0.0644 .
## as.numeric(self_rated_health)        -0.002236   0.007522  -0.297   0.7663
## as.numeric(age)                       0.009400   0.000609  15.433   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.547 on 20485 degrees of freedom
## Multiple R-squared:  0.01169,    Adjusted R-squared:  0.01154
## F-statistic: 80.75 on 3 and 20485 DF,  p-value: < 2.2e-16
```

Income ~ Self rated mental health + Age

```
##
## Call:
## lm(formula = as.numeric(income_family) ~ as.numeric(self_rated_mental_health) +
##     as.numeric(age), data = data_reformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7138 -1.3287 -0.2863  1.4317  2.9439
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.887478   0.042771  67.511   <2e-16 ***
## as.numeric(self_rated_mental_health) 0.012394   0.006651   1.864   0.0624 .
## as.numeric(age)                   0.009399   0.000609  15.433   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.547 on 20486 degrees of freedom
## Multiple R-squared:  0.01168,    Adjusted R-squared:  0.01159
## F-statistic: 121.1 on 2 and 20486 DF,  p-value: < 2.2e-16
```

Income ~ Age

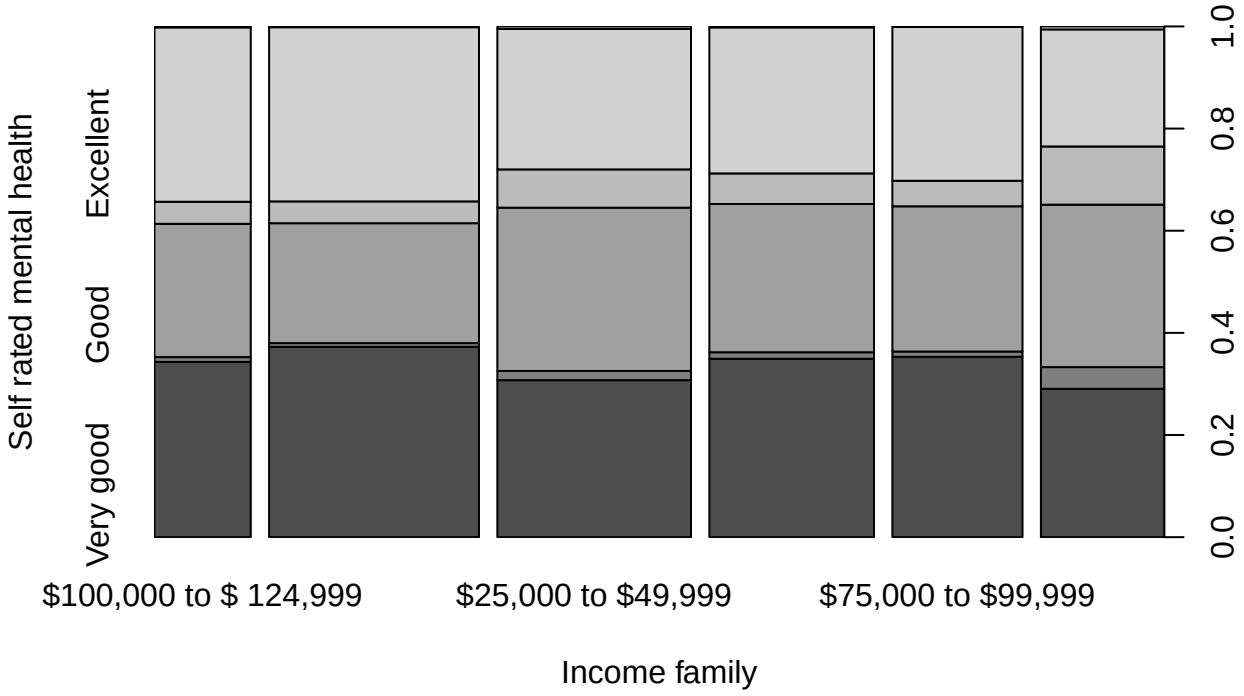
```
##
## Call:
## lm(formula = as.numeric(income_family) ~ as.numeric(age), data = data_reformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6896 -1.3292 -0.2878  1.4299  2.9220
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.936868   0.033571   87.48   <2e-16 ***
## as.numeric(age) 0.009409   0.000609   15.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.547 on 20487 degrees of freedom
## Multiple R-squared:  0.01152,    Adjusted R-squared:  0.01147
## F-statistic: 238.7 on 1 and 20487 DF,  p-value: < 2.2e-16
```

Income ~ Self rated mental health

```
##
## Call:
## lm(formula = as.numeric(income_family) ~ as.numeric(self_rated_mental_health),
##     data = income_mental)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.454 -1.414 -0.401  1.546  2.612
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        3.374440   0.029043 116.190   <2e-16 ***
## as.numeric(self_rated_mental_health) 0.013298   0.006688   1.988   0.0468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.556 on 20494 degrees of freedom
## Multiple R-squared:  0.0001928,  Adjusted R-squared:  0.0001441
## F-statistic: 3.953 on 1 and 20494 DF,  p-value: 0.0468
```

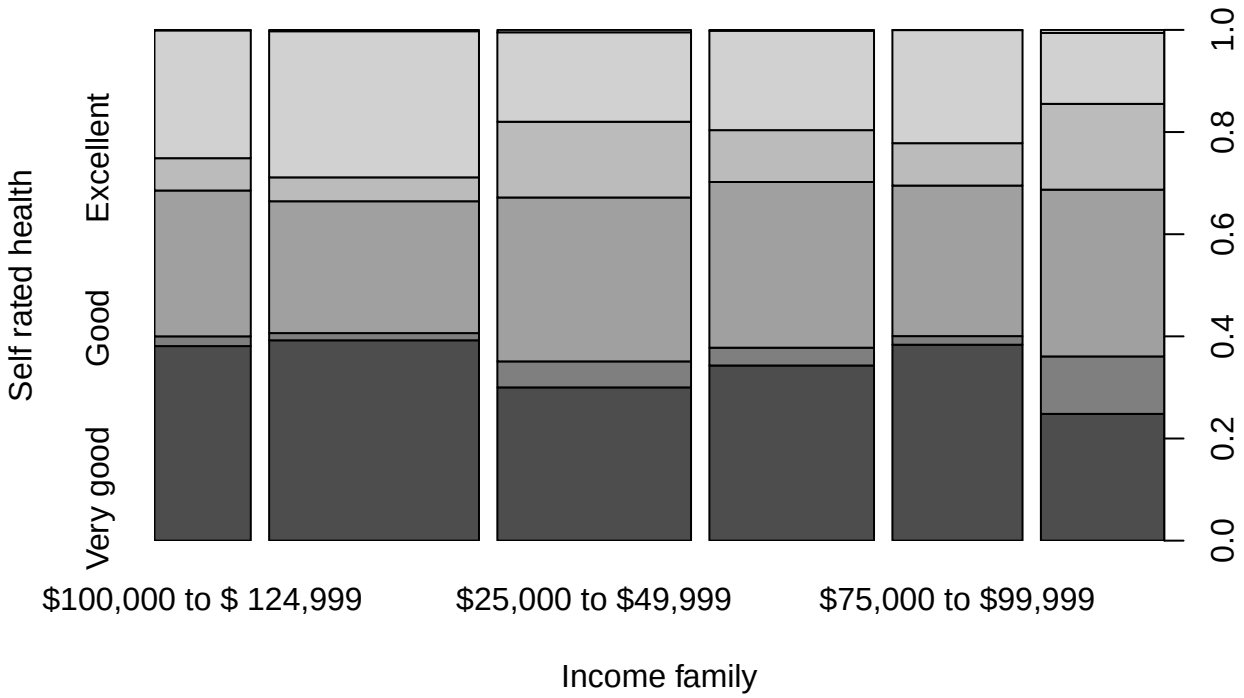
```r
attach(income_mental)
plot(income_family, self_rated_mental_health,
     main = "Figure2 Plot1 of Self rated mental health against Income family",
     xlab = "Income family", ylab = "Self rated mental health")
```

Figure2 Plot1 of Self rated mental health against Income family

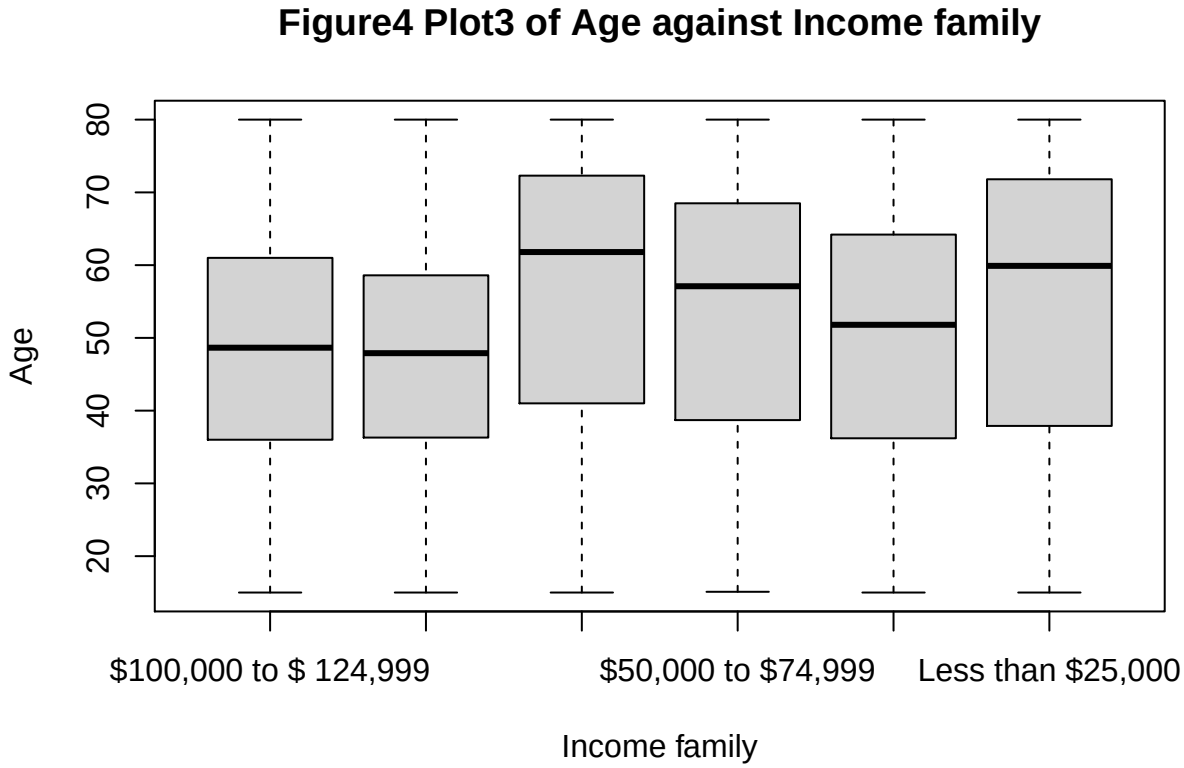



```
attach(data_reformed)
plot(income_family, self_rated_health,
     main = "Figure3 Plot2 of Self rated health against Income family",
     xlab = "Income family", ylab = "Self rated health")
```

Figure3 Plot2 of Self rated health against Income family



```r
plot(income_family, age,
     main = "Figure4 Plot3 of Age against Income family",
     xlab = "Income family", ylab = "Age")
```



The model I choose was the linear regression model. I fitted hour models here.
Mod 1 Income ~ Self rated mental health + Self Rated Health + Age

$\hat{Income} = 2.89 + 0.013self\_rated\_mental\_health - 0.002self\_rated\_health + 0.009age$

Mod 2 Income ~ Self rated mental health + Age

$\hat{Income} = 2.89 + 0.012self\_rated\_mental\_health - 0.009age$

Mod 3 Income ~ Age

$\hat{Income} = 2.94 + 0.009age$

Mod 4 Income ~ Self rated mental health

$\hat{Income} = 3.37 + 0.013self\_rated\_mental\_health$

Mod1 we fitted using all three variables, however, the p-value is extremely big, indicating that it is very unlikely for self rated health to be linearly correlated to Income, thus I dropped this variable and fitted another model Mod2 using Income with respect to self rated mental health and age. Mod3 and Mod4 is Income with respect to Self rated mental health and Age respectively to compare their p-value with Mod2.self rated mental health and age

Results

The first model Mod1 is a multiple linear regression model built to investigate if there is a linear relationship between income and self rated health, self rated mental health and age. From the summary table we could get: $\hat{Income} = 2.89 + 0.013self\_rated\_mental\_health - 0.002self\_rated\_health + 0.009age$, which means that one unit increase in self rated mental health, on average, income would increase 0.013 unit given all

other variables remain the same; one unit increase in self rated health, on average, income would decrease 0.002 unit given all other variables remain the same; one unit increase in age, on average, income would increase $ 0.009$ unit given all other variables remain the same.

The standard error of intercept is 0.047; the standard error of self rated mental health is 0.007; the standard error of self rated health is 0.008; while the standard error of age is 0.001. All the standard error are significant.

However, for the p-value, self rated mental health has a p-value: 0.064, the p-value for age is approximately 0.000, the p-values are not very significant, which supports H_a that income has correlation ship with self rated mental health and age. However, the p-value for self rated health is 0.766, which is significantly larger than 0.05, thus does not support H_a, as income is not likely to have a linear correlation ship with self rated health.

I noticed that the R^2 value for Mod1 is 0.01169, which is much smaller than 0.3, thus indicates that the linear relationship between income and the other three variables is very weak.

As the p-value for self rated health is very high, I dropped this variable and fitted other 3 multiple linear regression models: Mod2, Mod3 and Mod4.

For Mod2: $\hat{Income} = 2.89 + 0.012self\_rated\_mental\_health - 0.009age$. Thus we noticed that one unit increase in self rated mental health, on average, income would increase 0.012 unit given all other variables remain the same; while one unit increase in age, on average, income would increase $ 0.009$ unit given all other variables remain the same. It is very similar to Mod1.

According to summary table, The standard error of intercept is 0.042; the standard error of self rated mental health is $ 0.007$; while the standard error of age is 0.001. Again this is almost the same as Mod1 and is not significant as well.

The p-value for self rated mental health is 0.06 and the the p-value for age is again approximately 0.000. Though for self rated mental health is greater than 0.05, in Mod4 the p-value for it is below 0.05, thus I tend to conclude that it supports H_a that income has correlation ship with self rated mental health and age.

However, the R^2 for all Mods are less than 0.3, indicating that there is very weak linear relationship between income and self rated mental health and age as a whole and respectively as well.

From Figure 2, 3, 4 we could clearly witness that the income does not vary a lot as self rated mental health, self rated health and age changes.

Discussion

According to the result section, though the p-values for age and self rated mental health is low, the R^2 value for all four models are extremely low, which indicates that income does not have a linear relationship with self rated mental health, self rated health and age.Thus my hypothesis: the younger the respondent, the higher the individual rate on his/her health and metal health, the more this family might earn does not hold. Thus it is likely that income is not linearly correlated to self rated mental health, self rated health and age changes. This study shows that at least from the GSS 2017 data family income is not likely to be linearly affected by age self rated health and mental health, and could not act as an indicator or government to predict what population they could offer more help to.

Weaknesses

The data has some serious drawbacks, like I have discussed in Methodology section, a considerably amount of data was collected from elderly in their 80s, largely outweigh other ages. This could lead to bias and eventually result inaccuracy.

The study could address more factors, or fit another kind of model if linear regression model does not hold. Besides, during analysis, the p-value of self rated mental health in Mod1, 2 are all above 0.05, while in Mod4

it is below 0.05, the deviation might due to the addition of age to the model, but these two variable may be the main contributors leading to income changes from this study's perspective.

There are lots that could get improved, I could randomly pick some data from the section of people at age 80, this may make the analysis clearer and more accurate and could omit some serious age bias.

Next Steps

The linear regression model does not fit for the four variables that I have made investigation on, I wish to explore regarding income, and maybe we could fit another model like exponential and do some investigations. Besides that I should try to pick some data of from the 80 year old section and redo the linear regression model to if there is any differences to omit the severe bias caused by the people at age 80 largely outweigh other sections.

Being interested in income related issue, future steps could be we do a cluster sampling and pick 2 or 3 hospitals around the 10 provinces in Canada, and if the patients agree, we could collect the data from them, we could check if people who go to the hospital are usually people who have income above a certain level or do they have insurance covered, as I have heard people complaining about the high treatment fee in hospital.

References

1. Bruin, J. 2006. newtest: command to compute new test. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/stata/ado/analysis/.
2. https://stackoverflow.com/questions/32400916/convert-html-tables-to-r-data-frame 3.https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

Appendix

Github link: https://github.com/xuxiny17/A2.git

Figure4 Plot on self rated mental health

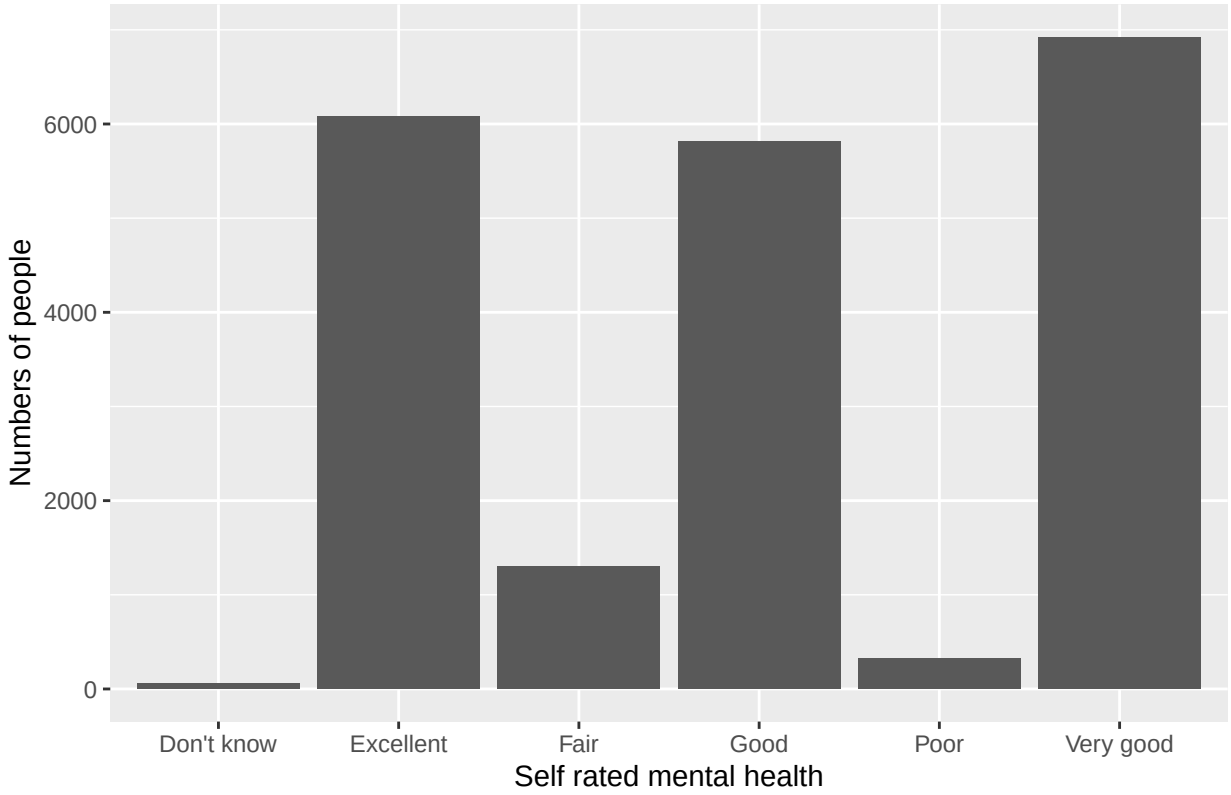
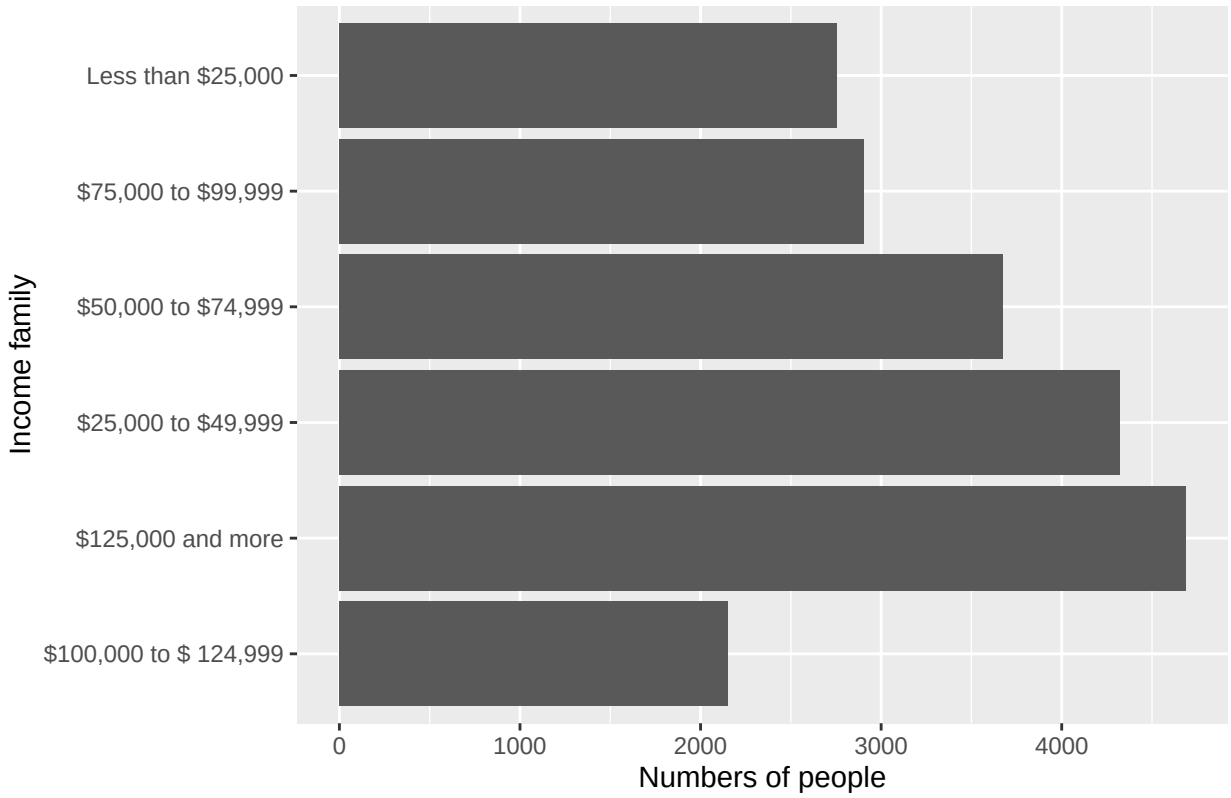
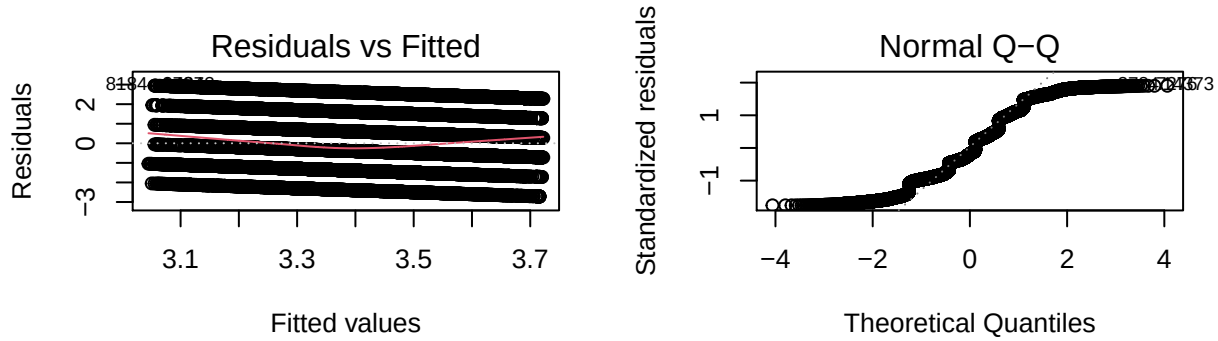
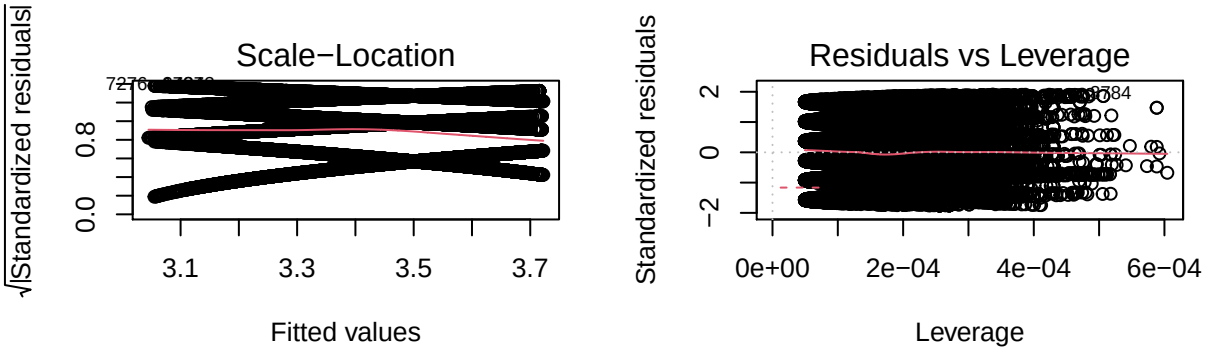
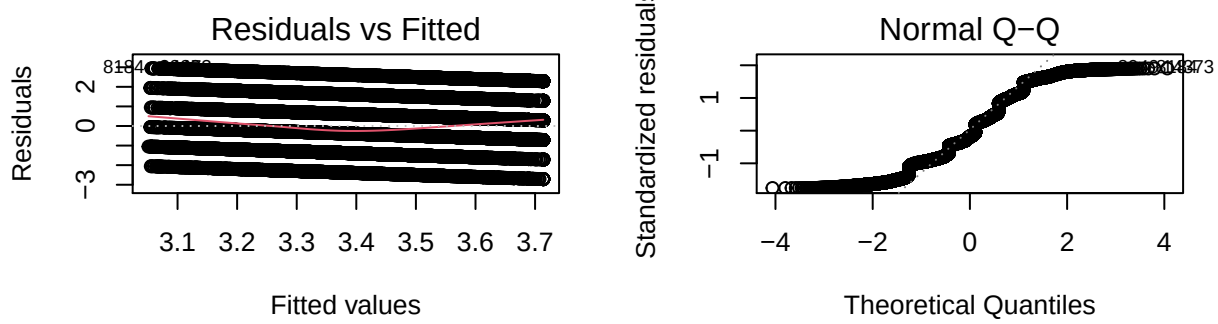


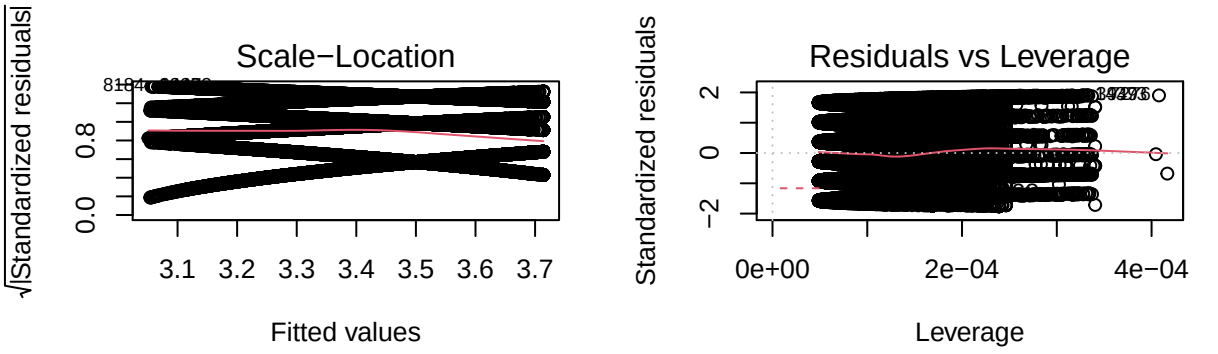
Figure5 Plot on numbers of people with different family income

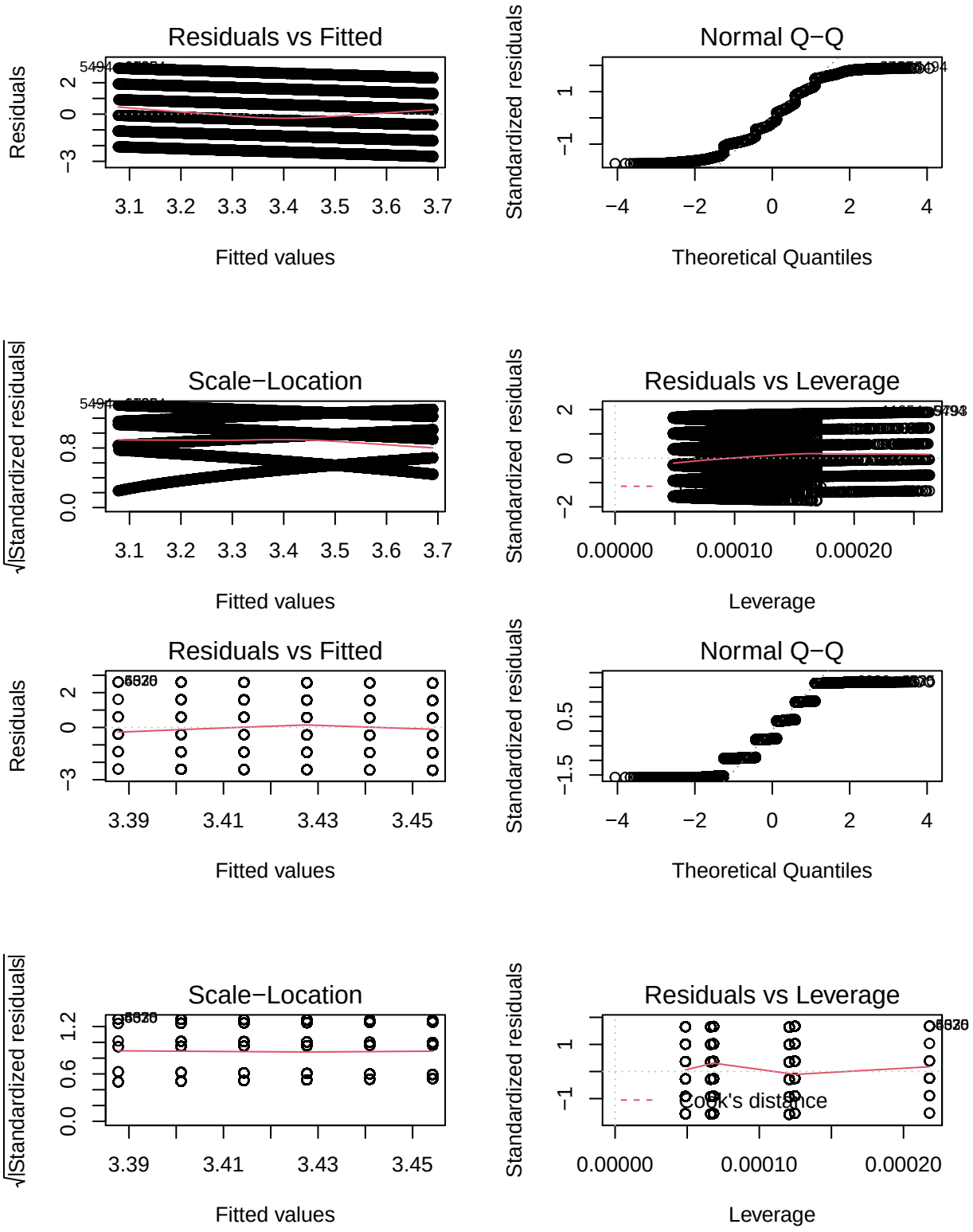



Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles


Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles


Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage




```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a factor
## response will be ignored

## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors

## Warning in abline(lm(income_family ~ self_rated_mental_health), col = "red"):
```

```
## only using the first two of 6 regression coefficients
```

Figure5 Plot4 of Self rated mental health against Income family