What Major Factors Affect Life Satisfaction?

Eung Kyu Kim, Dong Kyu Kim, Jiwon Chai

October 19th, 2020

What Major Factors Affect Life Satisfaction?

Names of Authors

Eung Kyu Kim (1003004278) Dong Kyu Kim (1003388675) Jiwon Chai (1005015396)

Date

October 19th, 2020

Abstract

The purpose of this study is to see if there is a direct relationship between an individual's income, family income, self-rated health,mental health,age,numbers of children and life satisfaction. We used three different regression methods which are simple linear regression, multiple linear regression, and logistic regression. Using simple linear regression, we have found that there is a weak linear relationship between households' income and feelings of lives. According to the multiple linear regression model, self-rated health factors showed some positive relationship to their feelings of lives. Lastly, under the logistic regression model, we found that the age and numbers of children have an impact on the level of happiness. The results have significant implications for determination of individuals' satisfaction of their lives.

Introduction

There are several possible factors that could affect one's life satisfaction. There is a famous saying "Money brings you happiness." However, there are several other factors that could also have an impact on the level of happiness/satisfaction such as, individuals' age, number of children, self-rated health. We believe that the amount of money that people earn is not the most important determination of true happiness.

First, we wanted to know how income affects life satisfaction. There were two variables about income: income_family, and income_respondent. We conducted two simple linear regression models which were 'income_family' vs 'feelings_life', and 'income_respondent' vs 'feelings_life' so we can know if an income is a crucial factor to make peoples' life happier.

Then, we conducted multiple linear regression model to find out if health is a crucial factor to make people happy. We are always told that health matters the most when it comes to life quality, and now we can check. There were two variables related to health: self_rated_health, and self_rated_mental_heatlh. This time, to see the overall relationship between health and life satisfaction, we used multiple linear regression. Lastly, models of logistic regression also known as logit model were conducted to find which variables affect the feelings of life the most.

Data

The data our group chose is the General Social Survey done on Canadian Families in 2017. The original data includes 20602 observations of the answers to 81 categories related to personal information. Before moving on to the analysis, observations that had any empty values were removed and 19832 observations were left. Out of the 81 variables, income, total children, self-rated health, and some other information have been selected for the analysis of the project. Out of the variables that were chosen, only age, the total number of children, and feelings of life were numerical values. Hence, it was difficult to analyze the data by linear regression. For certain analyses, categorical values had to be converted into a numerical value to be analyzed.

Model

For our data analysis, a simple linear regression model, multiple linear regression model and logit model were used to see the relationship between dependent variables and independent variables. To begin with, two simple linear regressions were made (Income_respondent vs feelings_life) and (Income_family vs feelings_life). This model gives an equation of Yˆhat = B0+B1ˆhat+ residuals ei. Yˆhat, which is the predicted value of Y in a regression where B0 is the intercept, which is the value of Y when dependent variable X=0. B1ˆhat is another predicted value of the amount of variable that changes when there is 1 unit change in X.

The linear regression model for income_respondent vs feelings_life gives the equation of Yˆhat = 7.622+0.143B1ˆhat +ei and the second simple linear regression model gives Yˆhat = 7.757+0.157B1ˆhat. Simple logics applied to multiple linear regression. Just one more variable was added. We have the MLR equation of Yˆhat= 4.802+ 0.304B1ˆhat + 0.574Bˆ2 hat+ei. The notation ei stands for error term in regression. The software used for modelling was R Studio.

Results

Two variables used are 'feelings_life' and 'income_family'. Setting income_family as an independent variable and feelings_life as a dependent variable, we drew the linear regression model on the graph. P-value was lower than 0.05, and the regression line had a weak positive relationship. Family income has a slight positive relationship between feelings of life. In the next simple linear regression, two variables used were 'income_respondent' and 'feelings_life'. P-value was still lower than 0.05 and it also showed a weak positive relationship.

Figure 9, the p-value of both self_rated_health and mental health are statistically significant, which means it has a strong evidence to reject the null-hypothesis (there are no relationship between health,mental health and feelings life).The estimate values of self_rated_health and self_rated_mental_health gives the multiple linear regression equation, which is Yˆhat = $0.304B0^{\text{hat}+0.57B1}$hat+ei, where ei is the residuals. Multiple R-squared value, 0.241 means approximately 24.1 % variation in feelings_life can be explained by self rated health and mental health.

We have an estimated slope for self rated health 0.304 or 95% confident that true slope is between 0.211 and 0.398. For mental health, we have an estimated slope of 0.57 and 95% confident that the true slope is between 0.471 and 0.676.

Figure 10, shows the relationship between self rated health, self rated mental health and feelings life are approximately linear, the variation looks constant. In Figure 11, the normal Q-Q line shows feelings life given self rated health and mental health are approximately normal.

Another method used to find the answer to the original goal of the study is the logistic regression function. Logistic regression function as known as the logit model is used when the outcome is binary and when we want to know how each variable affects the outcome. In this analysis, two models of logistic regression functions were drawn each comparing different variables.

The first logit model was between the variables of age and income of the respondents and how each variable affects the probability of being happy. The outcome of the model is shown in Figure 6. From the figure, it can be seen that the slope of the graph is negative, so it can be concluded that as the age increases, the predicted probability of being happy will decrease. Additionally, the different colours in the graph indicate the different categories of income of the respondents.

Another logit model is shown in Figure 8 and this model was between the variables of self-rated health and the total number of children. This graph shows that as the respondents self-rate their health highly, the predicted probability of being happy increases. Also, as the number of children increases, the probability of being happy increases.

Discussion

The original goal of the study was to find which variables affect people feeling happy with life the most. Some people say money can buy us happiness. We wanted to know if money is a good indicator to evaluate one's happiness. The original data had more than 20,000 rows of data, so we randomly sampled using the sample() function, and created a new data frame with only 1000 rows of data. As shown in results, R-squared values for both simple linear regression cases were lower than expected. We can conclude that income level could somehow affect the happiness level, but cannot be the major factor.

The p-values that were found in the MLR model were almost equal to 0, which means there is strong evidence that rejects my null hypothesis (there is no relationship between self- rated health, mental health and life-feelings). However, the adjusted squared R-value, which is an indication of how well terms fit the curve. In this study, the low value of adjusted R-squared tells us that our chosen variable might not be appropriate combinations or we have missed an important predicted variable. However, studies have shown that any field that is related to human behavior such as psychology has R-squared value lower than 50% (Minitab., 2013). It is shown from the plots drawn from the logit models show that as income increases, the predicted probability of being happy also increases. Also, an important component to focus on the second logit model is the fact that as the respondents rate their health highly, the number of children hardly affects the probability of being happy.

Simple linear regression model showed that income level has a slight positive relationship with life satisfaction. Multiple linear regression model showed that the self-rated health status also has a slight positive relationship with life satisfaction. Logistic regression model stated that the younger (suppose healthier) and the richer you are, the happier you will be. Looking at all those three models, we can conclude that both income level and health status were somehow affecting life satisfaction, but the degree of the impact were too small to say that they are the major factors. Seeing those data, we realized that there is no absolute factor that could make someone especially happier than the other. People will be able to have high life satisfaction no matter how rich you are or how healthy you are.

Weaknesses

One of the weaknesses of the study and the analysis done in this report is the lack of numerical values in the dataset. In the dataset used, nearly all the values of the variables were categorical variables instead of numerical variables. For example, one of the important variables used during the analysis was the income of the respondent and the family. However, the values were given as categorical values such as "$25,000 to $49,999", which is a wide range.

Since the range was wide for each category, people were only divided into 5 categories depending on their income. Due to the small number of categories and non-numerical value, it was difficult to accurately find how income affects feelings of happiness. Therefore, the analysis and the study would have given better results if the values of the variables were given as numerical values or have more categories for each variable which has a narrow range of values.

In addition to limited variation, the comparison between the dependent factors were not discussed in depth. As our paper focused on the what are the most important factors that affect people' happiness. We have investigated some possible factors but did not have a specific comparison between the variables. For example, determining the rank of dependent variables that affect the independent variable by assigning the number to their rank could be one of the possible improvements. This is one another thing that could be improved subsequently.

Next Steps

Although the report has been done, there could always be improvements made. One of the improvements that could be made is using a different analysis method with the data chosen. In this project, methods of SLR, MLR and Logit were used, but a different method such as Bayesian or Hierarchical Model could give better results for this project. Another improvement that can help this project is conducting a follow-up survey that can help with specification of values. As it has been stated earlier, most values of variables in the data frame are categorical values which limit us from doing various analyses. Hence, if another survey is conducted asking for numerical values for the variables, the analysis will be significantly improved.

References

Faculty of Arts & Sciences University of Toronto. (2017). Data Centre. http://dc.chass.utoronto.ca/myaccess.html. Institute for Digital Research & Education Statistical Consulting, U. C. L. A. (2013, December 16). LOGIT REGRESSION | R DATA ANALYSIS EXAMPLES. Introduction to SAS. https://stats.idre.ucla.edu/r/dae/logit-regression/. Minitab Blog. (2013, May 30). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.

Appendix

Codes Written for Data

```
ps2_data <- gss

ps2_data <- ps2_data %>%
  select(caseid, age, feelings_life, self_rated_health, self_rated_mental_health,
         income_family, income_respondent, average_hours_worked, total_children,
         total_children) %>%

  mutate(self_rated_health = case_when(
    self_rated_health=="Poor" ~ as.numeric(1),
    self_rated_health=="Fair" ~ as.numeric(2),
    self_rated_health=="Good" ~ as.numeric(3),
    self_rated_health=="Very good" ~ as.numeric(4),
    self_rated_health=="Excellent" ~ as.numeric(5)

  ))  %>%

  mutate(self_rated_mental_health = case_when(
    self_rated_mental_health=="Poor" ~ as.numeric(1),
    self_rated_mental_health=="Fair" ~ as.numeric(2),
```

```
    self_rated_mental_health=="Good" ~ as.numeric(3),
    self_rated_mental_health=="Very good" ~ as.numeric(4),
    self_rated_mental_health=="Excellent" ~ as.numeric(5)

  )) %>%

  mutate(income_family = case_when(
    income_family=="Less than $25,000" ~ as.numeric(1),
    income_family=="$25,000 to $49,999" ~ as.numeric(2),
    income_family=="$50,000 to $74,999" ~ as.numeric(3),
    income_family=="$75,000 to $99,999" ~ as.numeric(4),
    income_family=="$100,000 to $ 124,999" ~ as.numeric(5),
    income_family=="$125,000 and more" ~ as.numeric(6)

  )) %>%

  mutate(income_respondent = case_when(
    income_respondent=="Less than $25,000" ~ as.numeric(1),
    income_respondent=="$25,000 to $49,999" ~ as.numeric(2),
    income_respondent=="$50,000 to $74,999" ~ as.numeric(3),
    income_respondent=="$75,000 to $99,999" ~ as.numeric(4),
    income_respondent=="$100,000 to $ 124,999" ~ as.numeric(5),
    income_respondent=="$125,000 and more" ~ as.numeric(6)

  )) %>%

  mutate(average_hours_worked = case_when(
    average_hours_worked=="NA" ~ as.numeric(1),
    average_hours_worked=="0.1 to 29.9 hours" ~ as.numeric(2),
    average_hours_worked=="30.0 to 40.0 hours" ~ as.numeric(3),
    average_hours_worked=="40.1 to 50.0 hours" ~ as.numeric(4),
    average_hours_worked=="50.1 hours and more" ~ as.numeric(5)
  ))

#Omitting rows that ha 'na' in the columns

ps2_data_complete <- na.omit(ps2_data)
```

SLR Model 1: Income Family

```
#Randomly selecting observations from the data frame

random_select_rows <- ps2_data_complete[sample(nrow(ps2_data_complete), 1000), ]
SLR_model_incomefamily <- lm(formula = feelings_life ~ income_family, data = random_select_rows)
```

```
summary(SLR_model_incomefamily)
```

```
##
## Call:
## lm(formula = feelings_life ~ income_family, data = random_select_rows)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8194 -0.8194  0.1806  1.1806  2.3673
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.44605    0.11011  67.622  < 2e-16 ***
## income_family  0.18665    0.02797   6.674 4.11e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.551 on 998 degrees of freedom
## Multiple R-squared:  0.04273,    Adjusted R-squared:  0.04177
## F-statistic: 44.55 on 1 and 998 DF,  p-value: 4.114e-11
```

Figure 1. Summary of Family Income SLR

```
ggplot(data = random_select_rows, ) +
  aes(x = income_family, y = feelings_life) +
  geom_point() +
  theme(text = element_text(size=20)) +
  ggtitle("Family income vs Feelings of life") +
  theme(text = element_text(size = 13)) +
  labs(y = "Feelings of Life", x= "Family's Income") +
  geom_smooth(method='lm', formula = y~x, se = FALSE)
```

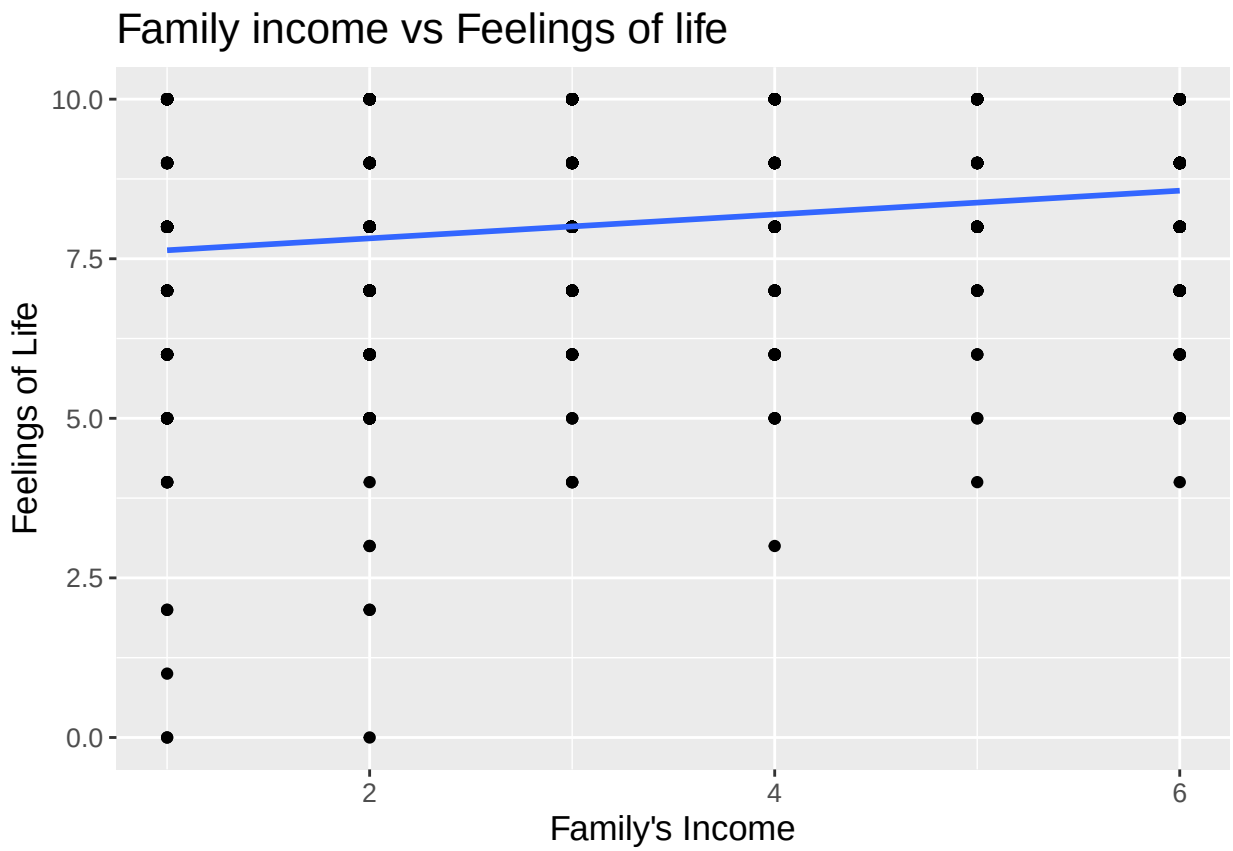


Figure 2. Plot of Family Income vs Feelings of Life

SLR Model 2: Income Respondent

```
SLR_model_income_respondent <- lm(formula = feelings_life ~ income_respondent, data = random_select_rows
```

```
summary(SLR_model_income_respondent)
```

```
##
## Call:
## lm(formula = feelings_life ~ income_respondent, data = random_select_rows)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0557 -0.8927  0.1073  1.1073  2.1073
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.72962    0.10034  77.034  < 2e-16 ***
## income_respondent  0.16306    0.03797   4.295 1.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.571 on 998 degrees of freedom
## Multiple R-squared:  0.01814,    Adjusted R-squared:  0.01716
## F-statistic: 18.44 on 1 and 998 DF,  p-value: 1.922e-05
```

Figure 3. Summary of Respondent Income SLR

```
ggplot(data = random_select_rows, ) +
  aes(x = income_respondent, y = feelings_life) +
  geom_point() +
  theme(text = element_text(size=20)) +
  ggtitle("Income vs feelings life") +
  theme(text = element_text(size = 13)) +
  labs(y = "feelings life", x= "Individual Income") +
  geom_smooth(method='lm', formula = y~x, se = FALSE)
```

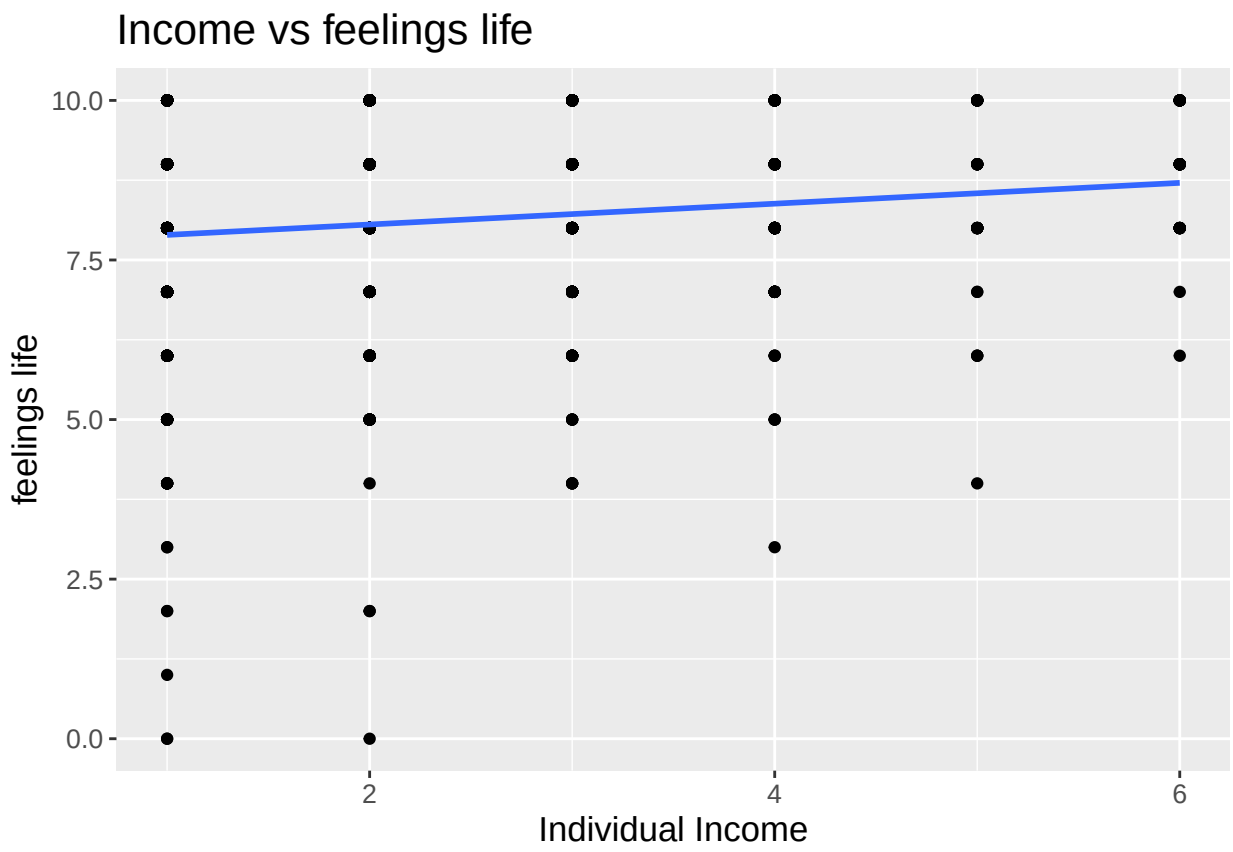


Figure 4. Plot of Feelings of Life vs Respondent Income

Logit Model 1:

```
ps2_ex <-ps2_data %>%
```

```
  select(age, feelings_life, self_rated_health, self_rated_mental_health,
         income_family, income_respondent, average_hours_worked, total_children,
         total_children) %>%

  mutate(feelings_life = case_when(
    feelings_life=="1" ~ as.numeric(0),
    feelings_life=="2" ~ as.numeric(0),
    feelings_life=="3" ~ as.numeric(0),
    feelings_life=="4" ~ as.numeric(0),
    feelings_life=="5" ~ as.numeric(0),
    feelings_life=="6" ~ as.numeric(1),
    feelings_life=="7" ~ as.numeric(1),
    feelings_life=="8" ~ as.numeric(1),
    feelings_life=="9" ~ as.numeric(1),
    feelings_life=="10" ~ as.numeric(1)

  ))

ps2_ex_complete_logit1 <- na.omit(ps2_ex)


ps2_ex_complete_logit1$income_respondent<- factor(ps2_ex_complete_logit1$income_respondent)

mylogit <- glm(feelings_life ~ age + total_children + income_respondent, data = ps2_ex_complete_logit1,
```

```
summary(mylogit)
```

```
##
## Call:
## glm(formula = feelings_life ~ age + total_children + income_respondent,
##     family = "binomial", data = ps2_ex_complete_logit1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8967   0.2998   0.3780   0.4453   0.5500
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.376846   0.087920  27.034  < 2e-16 ***
## age                -0.007057   0.001681  -4.198 2.70e-05 ***
## total_children      0.076236   0.021037   3.624  0.00029 ***
## income_respondent2  0.393105   0.064067   6.136 8.47e-10 ***
## income_respondent3  0.796556   0.084543   9.422  < 2e-16 ***
## income_respondent4  1.015314   0.121231   8.375  < 2e-16 ***
## income_respondent5  1.278704   0.203461   6.285 3.28e-10 ***
## income_respondent6  1.710862   0.241719   7.078 1.46e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10310  on 19761  degrees of freedom
## Residual deviance: 10060  on 19754  degrees of freedom
```

```
## AIC: 10076
##
## Number of Fisher Scoring iterations: 6
```

Figure 5. Summary of First Logit Model

```
newdata1 <- with(ps2_ex_complete_logit1, data.frame(age= mean(age), total_children = mean(total_childre

newdata1$rankP <- predict(mylogit, newdata = newdata1, type = "response")

newdata11 <- with(ps2_ex_complete_logit1, data.frame(age = rep(seq(from = 1, to = 100, length.out= 100)
                                                     total_children = mean(total_children),
                                                     income_respondent = factor(rep(1:5, each = 100))))



newdata111 <- cbind(newdata11, predict(mylogit, newdata = newdata11, type = "link",
                                       se = TRUE))
newdata111 <- within(newdata111, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})

#head(newdata111)
```

```
ggplot(newdata111, aes(x = age, y = PredictedProb)) +
        geom_ribbon (aes(ymin = LL,
                         ymax = UL, fill = income_respondent), alpha = 0.2) +
        geom_line(aes(colour = income_respondent),
                  size = 1) +
      xlab("Age of Respondents") +
      ylab("Predicted Probability of Being Happy") +
      ggtitle("Plot of Probabilities of Being Happy Considering the
              Differences of Income and Age") +
      scale_fill_discrete(name = "Income of Respondents" , labels = c("Less than $25,000", "$25,000 to
```

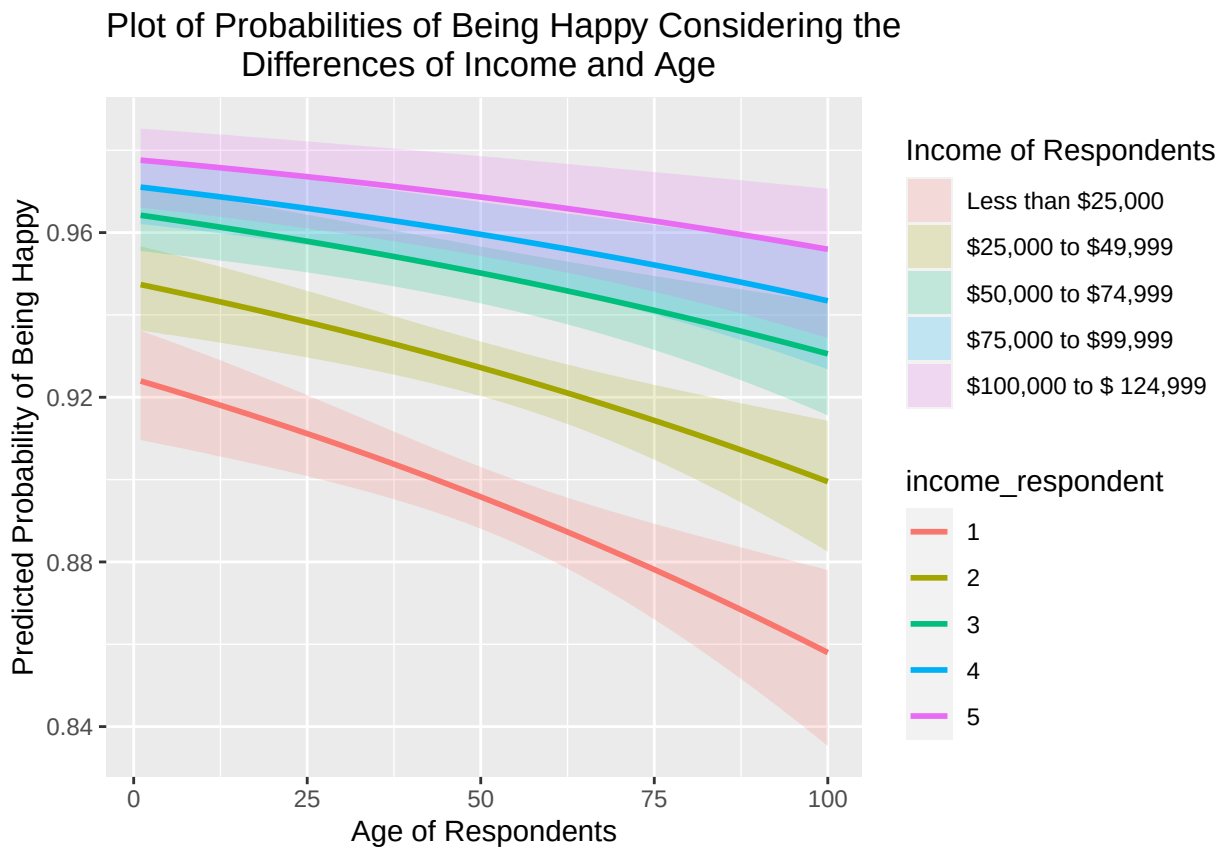


Figure 6. Plot of Probabilities of Being Happy Considering the Differences of Income and Age

Logit Model 2:

```
ps2_ex_complete_logit2 <- na.omit(ps2_ex)


ps2_ex_complete_logit2$self_rated_health<- factor(ps2_ex_complete_logit2$self_rated_health)

mylogit2 <- glm(feelings_life ~ age + total_children + self_rated_health, data = ps2_ex_complete_logit2
```

```
summary(mylogit2)
```

```
##
## Call:
## glm(formula = feelings_life ~ age + total_children + self_rated_health,
##     family = "binomial", data = ps2_ex_complete_logit2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9597   0.2132   0.2546   0.3931   1.2141
##
```

```
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.249769   0.128566  -1.943    0.052 .
## age                 0.007884   0.001864   4.231 2.33e-05 ***
## total_children      0.086386   0.021609   3.998 6.40e-05 ***
## self_rated_health2  1.044995   0.095116  10.987  < 2e-16 ***
## self_rated_health3  2.165586   0.091901  23.564  < 2e-16 ***
## self_rated_health4  3.208622   0.106560  30.111  < 2e-16 ***
## self_rated_health5  3.620560   0.134798  26.859  < 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: 10310.4  on 19761  degrees of freedom
## Residual deviance:  8859.4  on 19755  degrees of freedom
## AIC: 8873.4
##
## Number of Fisher Scoring iterations: 6
```

Figure 7. Summary of Second Logit Model

```
newdata2 <- with(ps2_ex_complete_logit2, data.frame(age= mean(age), total_children = mean(total_childre

newdata2$rankP <- predict(mylogit2, newdata = newdata2, type = "response")



newdata22 <- with(ps2_ex_complete_logit2, data.frame(total_children = rep(seq(from = 0, to = 9, length.
                                                     age = mean(age),
                                                     self_rated_health = factor(rep(1:5, each = 100))))



newdata222 <- cbind(newdata22, predict(mylogit2, newdata = newdata22, type = "link",
                                       se = TRUE))
newdata222 <- within(newdata222, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})

#head(newdata222)
```

```
ggplot(newdata222, aes(x = total_children, y = PredictedProb)) + geom_ribbon (aes(ymin = LL,
                                                                                  ymax = UL, fill = self_rated_heal
        geom_line(aes(colour = self_rated_health),
                  size = 1) +
      xlab("Total Number of Children") +
      ylab("Predicted Probability of Being Happy") +
      ggtitle("Plot of Probabilities of Being Happy Considering the
              Differences of Self Rated Health and Total Number of Children ")
```

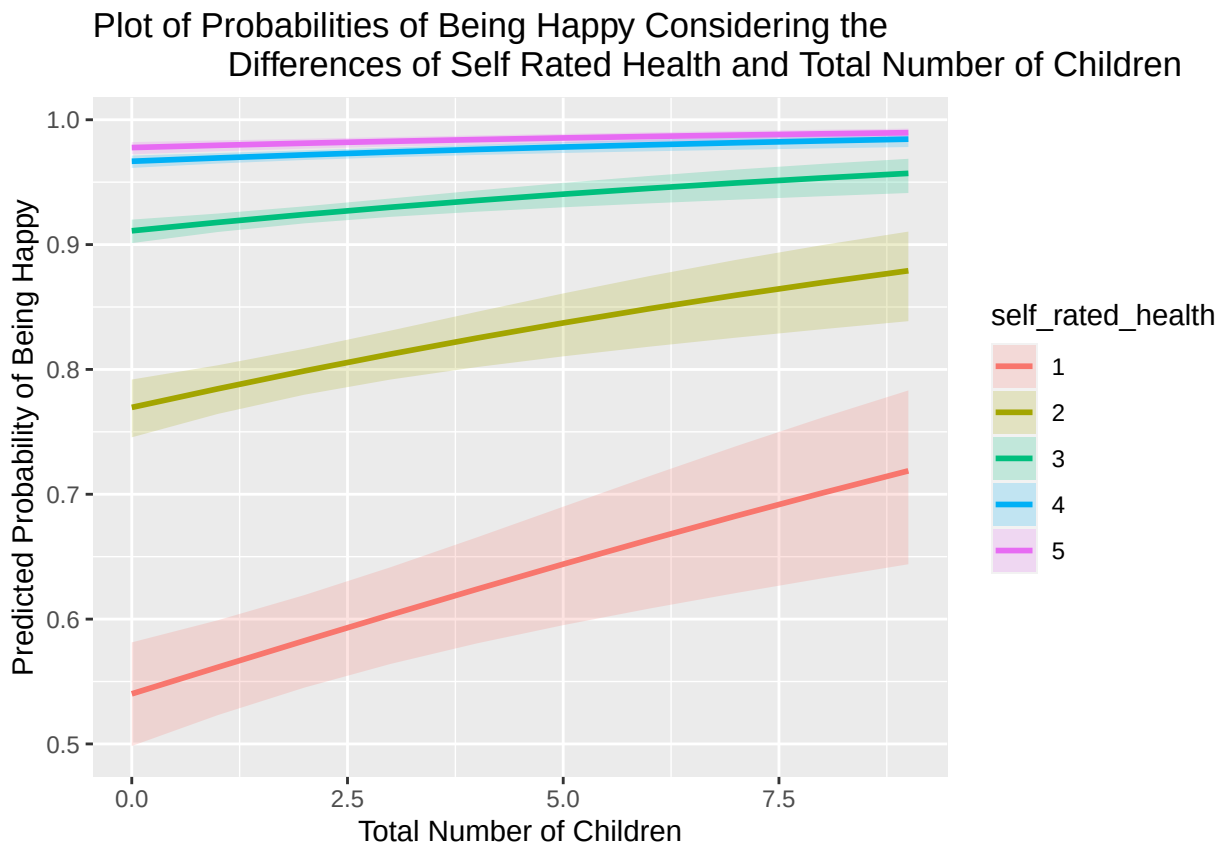


Figure 8. Plot of Probabilities of Being Happy Considering the Differences of Self Rated Health and Total Number of Children

Multiple linear regression line

```
MLR_model_health <- lm(formula=feelings_life~self_rated_health+self_rated_mental_health, data= random_s
```

```
summary (MLR_model_health)
```

```
##
## Call:
## lm(formula = feelings_life ~ self_rated_health + self_rated_mental_health,
##     data = random_select_rows)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7016 -0.7701  0.0596  0.8835  3.7955
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.70746    0.19968  23.575  < 2e-16 ***
## self_rated_health         0.32683    0.04605   7.097 2.42e-12 ***
```

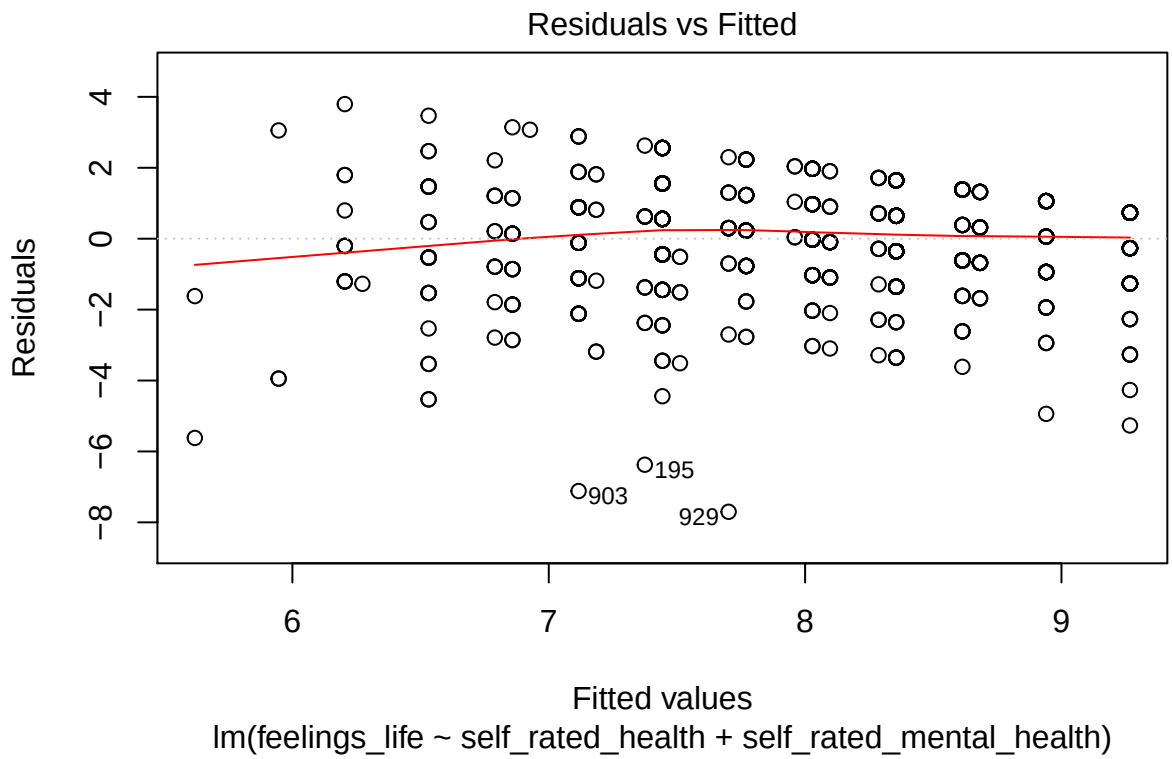
```
## self_rated_mental_health  0.58511    0.05135  11.395  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.389 on 997 degrees of freedom
## Multiple R-squared:  0.2337, Adjusted R-squared:  0.2321
## F-statistic:   152 on 2 and 997 DF,  p-value: < 2.2e-16
```

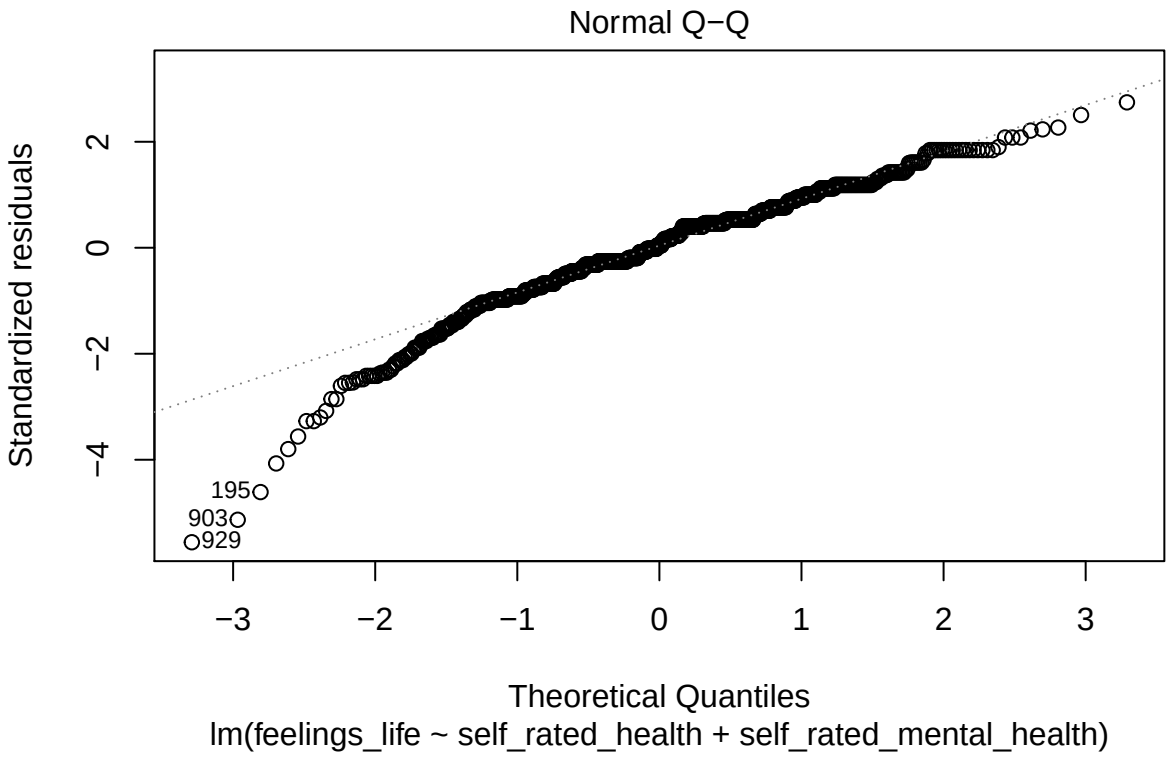
Figure 9. MLR Summary of Feelings Life Considering Self-Rated Health

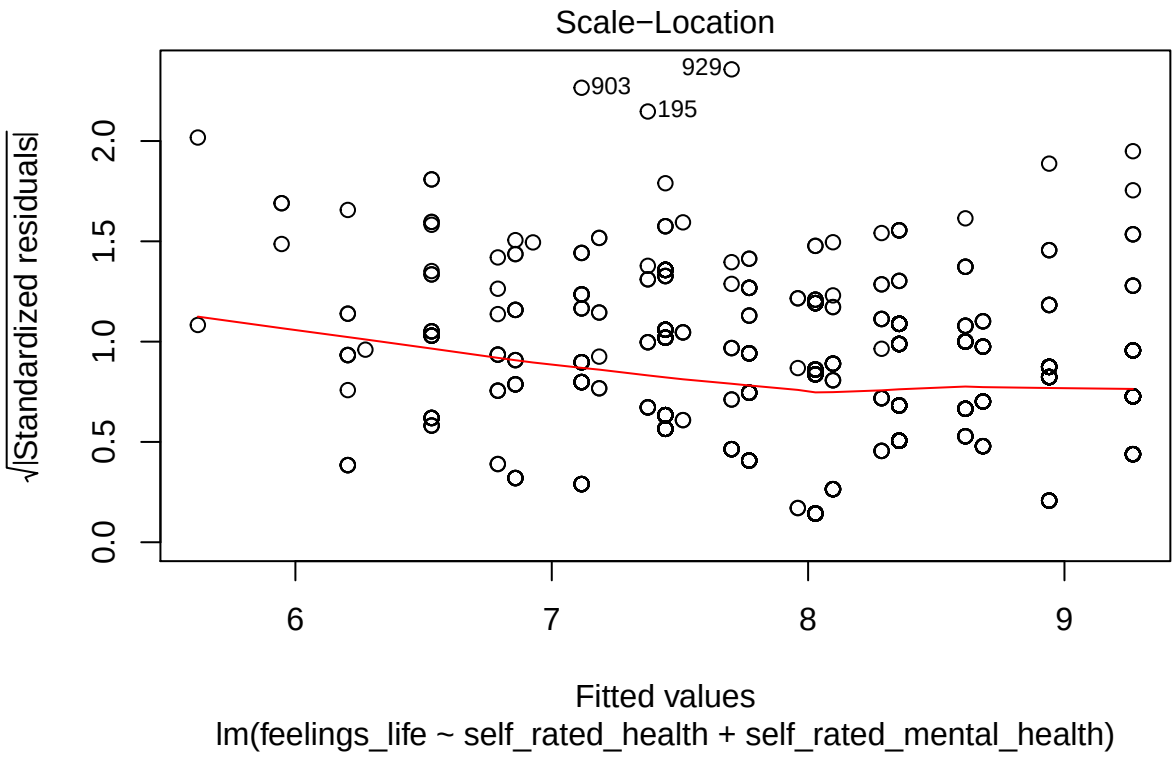
```
#Confidence intervals for the model coefficients
confint(MLR_model_health, conf.level = 0.95)
```

```
##                               2.5 %    97.5 %
## (Intercept)              4.3156136 5.0993078
## self_rated_health        0.2364610 0.4171988
## self_rated_mental_health 0.4843522 0.6858776
```

```
plot(MLR_model_health)
```




Normal Q−Q
Standardized residuals
2
0
−2
−4
195
903
929
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(feelings_life ~ self_rated_health + self_rated_mental_health)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
903
929
195
6
7
8
9
Fitted values
lm(feelings_life ~ self_rated_health + self_rated_mental_health)

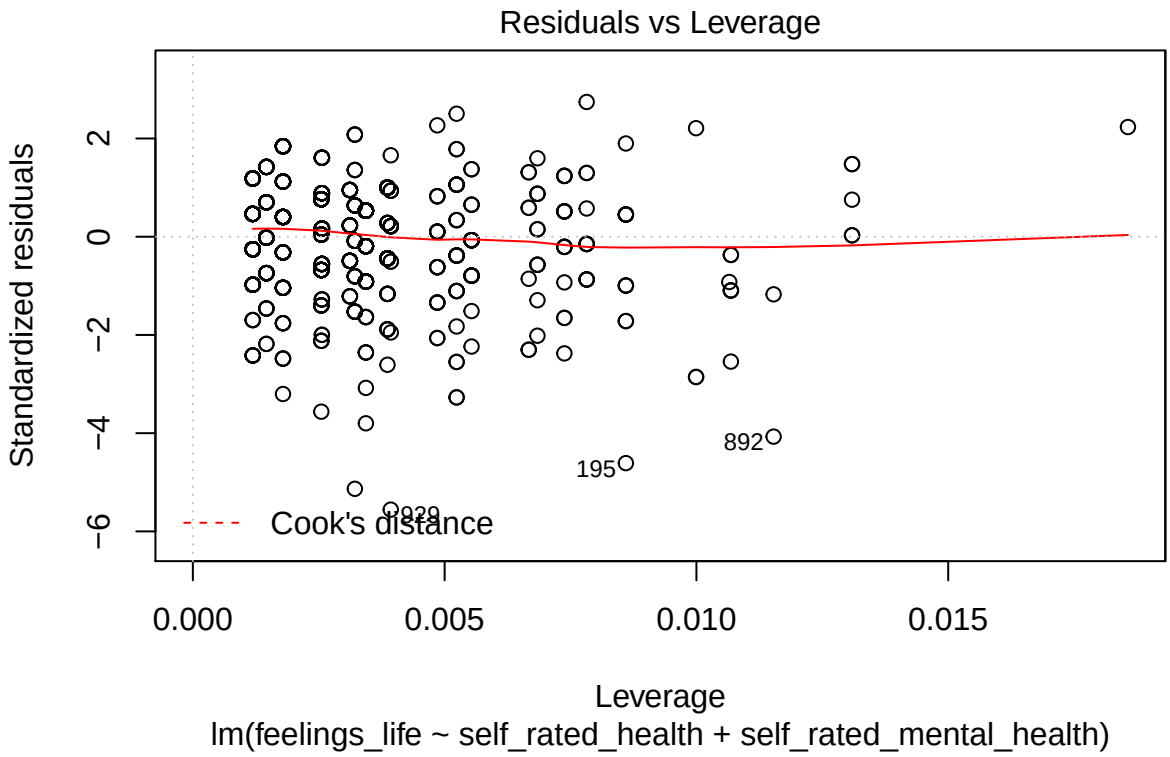


Figure 10. Residual vs Fitted of MLR Plot

Figure 11. Normal Q-Q of MLR Plot

Figure 12. Scale-Location of MLR Plot