MATH 6357
Linear Models and Design of Experiment
Final Group Project Fall 2020

Instructor: Wenshuang Wang

Which significant predictors most likely affect human life expectancy?

Authors:

Thuy Le

Sara Nafaryeh

Tony Nguyen

Thomas Su

Table of Contents

Introduction:

The World Happiness Report is a landmark survey of the state of global happiness, ranking 153 countries by how happy their citizens perceive themselves to be in that country. The scores and rankings make use of information from the Gallup World Survey.

In our report, we will be exploring 6 selected feature predictors, out of the total 18 in the data, that we believe to have an effect on our selected-response variable. We are interested in the relationship between life expectancy and other variables such as GDP, social support, ladder score, freedom to make life choices, perception of corruption, and generosity. For this report, we focus on answering the **research question**: "How are the predictors-logged of GDP, ladder score, social support, freedom to make life choices, perception of corruption, generosity - affect the Life Expectancy?"

Response (Y)
Life Expectancy: average number of years that a newborn can expect to live in "full health"—in other words, not hampered by disabling illnesses or injuries.

Predictors (X1, X2, X3, X4, X5, X6)

(X1) = logged of GDP: natural log of GDP. GDP is calculated in purchasing power parity (PPP) at constant 2011 international dollar prices from the November 28, 2019 update of the World Development Indicators (WDI) and is a metric that breaks down a country's economic output per person and is calculated by dividing the GDP of a country by its population.

(X2) Ladder score: Happiness score or subjective well-being. The national average response to the question of life evaluations. The English wording of the question is "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"

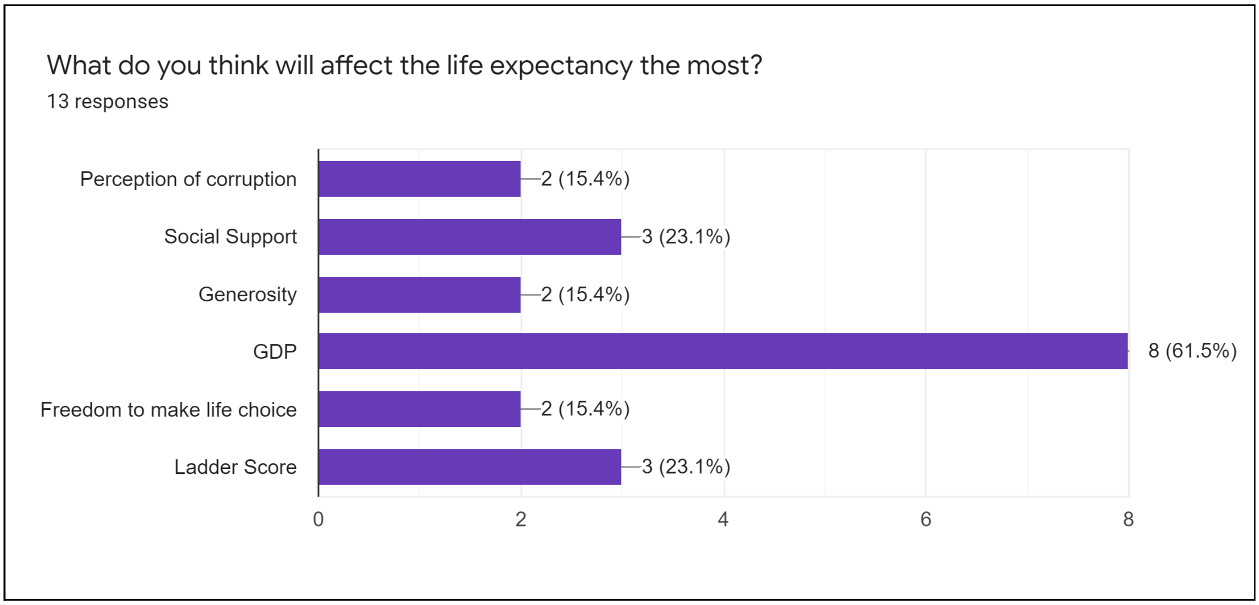
(X3) = Social Support: Having someone to count on in times of trouble. Is the national average of the binary responses (either 0 or 1) to the GWP question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

(X4) = Freedom to make choices: Freedom to make life choices is the national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?**" (**in percentage)

(X5) = Perception of Corruption: The measure is the national average of the survey responses to two questions in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?" The overall perception is just the average of the two 0-or-1 responses. In case the perception of government corruption is missing, we use the perception of business corruption as the overall perception. The corruption perception at the national level is just the average response of the overall perception at the individual level.

(X6) = Generosity: Generosity is the residual of regressing the national average in response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.

Expectation - Hypothesis: Before presenting our report to the class, we created a survey and decided to ask our classmates. Based on their subjective opinions, we asked them which feature they believed to have the most significance to predict **(Y)** Life Expectancy without any computation in data analysis. Out of a total of 13 responses, (**X1**) logged GDP took the lead at 8 votes (61.5%) , followed by a tie of 3 (23.1%) votes for (**X2**) Ladder score, (**X3**) Social Support.



Methodology:

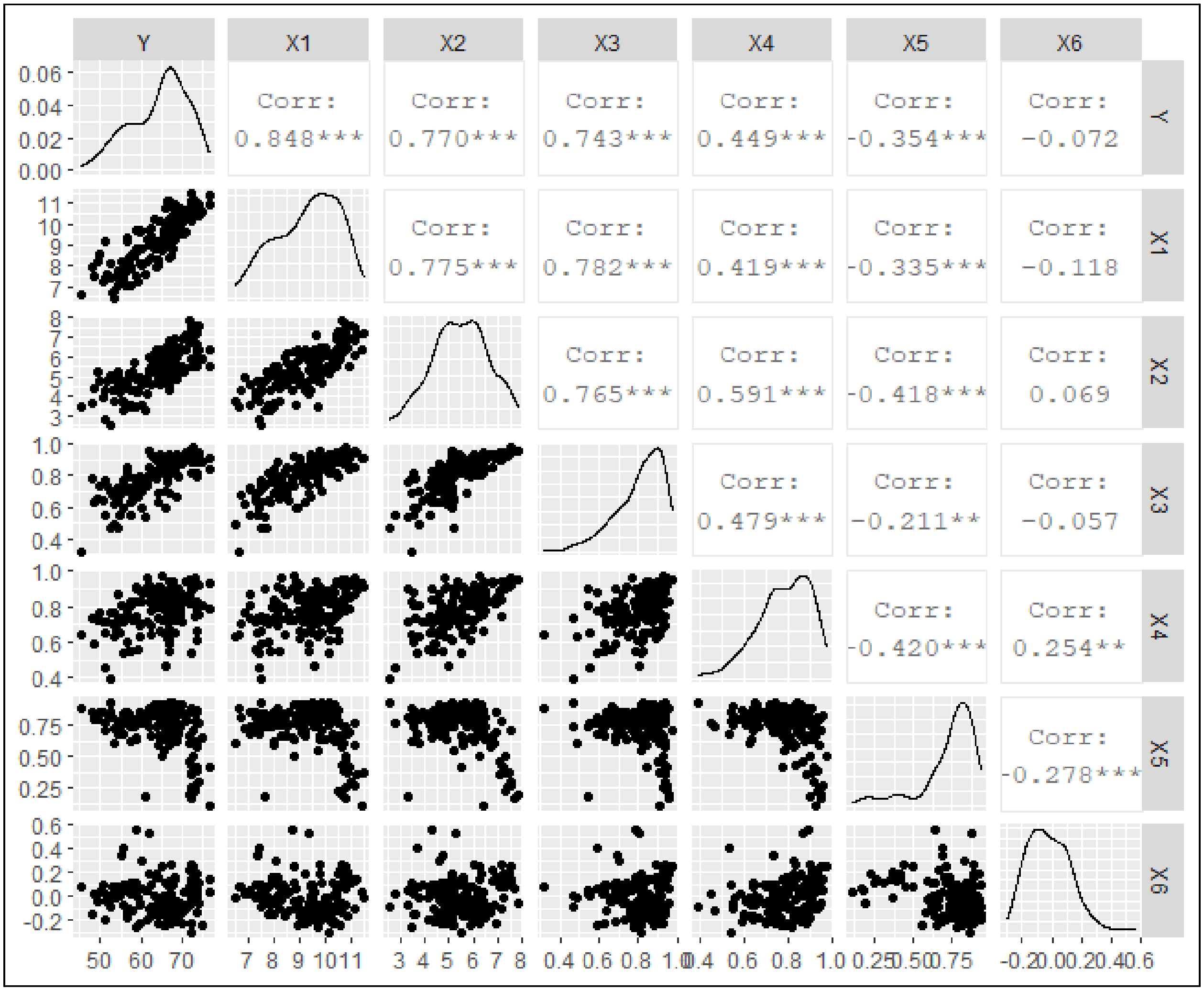
We decided to examine the data with three methods: scatter plot, stepwise, and hypothesis testing. With the scatter plot method, we use the ggpairs function in R to display the distribution of data between the response variable (Life expectancy) and the predictor variables (logged of GDP, ladder score, social support, freedom to make choice, corruption, and generosity). This function also shows the correlation between the variables which gives us an overall observation of the variables' relationship. With this beginning look, we will have our first predictions for the data set.

In order to decide the best subset of variables that have a significant impact on life expectancy, we conduct the Stepwise in R, with both "forward" and "backward" directions. The main idea of this function is to choose the subset of variables that have the lowest AIC. For the "forward" direction, the function starts with the smallest subset and builds up the subset to the full model by adding the variables to the model. On the other side, we start with the full model with all six predictors and drop each predictor until we can find the most substantial subset in the "backward" direction. We eventually choose to display our analysis with the "backward" direction as we want to provide the full glance for the model first before finding the fittest model for the data set.

After having the subset, we perform the Hypothesis test to reassure the decision one more time if the Stepwise testing is correct. Our intention is doing 2 tests: a full model with the null hypothesis testing for all the coefficients to be zero vs full model with the null hypothesis testing for only the coefficients of the variables that get eliminated in the Stepwise function to be zero.
If the hypothesis test matches the Stepwise, we are more confident to conclude our decision for the data set.

Data Analysis:

Using the function ggpairs() in the R package ggplot2, we are able to display both the correlation matrix and the scatter plot of the data in one nice table. It appears that the features X1, X2, and X3 show signs of linearity between themselves and our Y predictor.

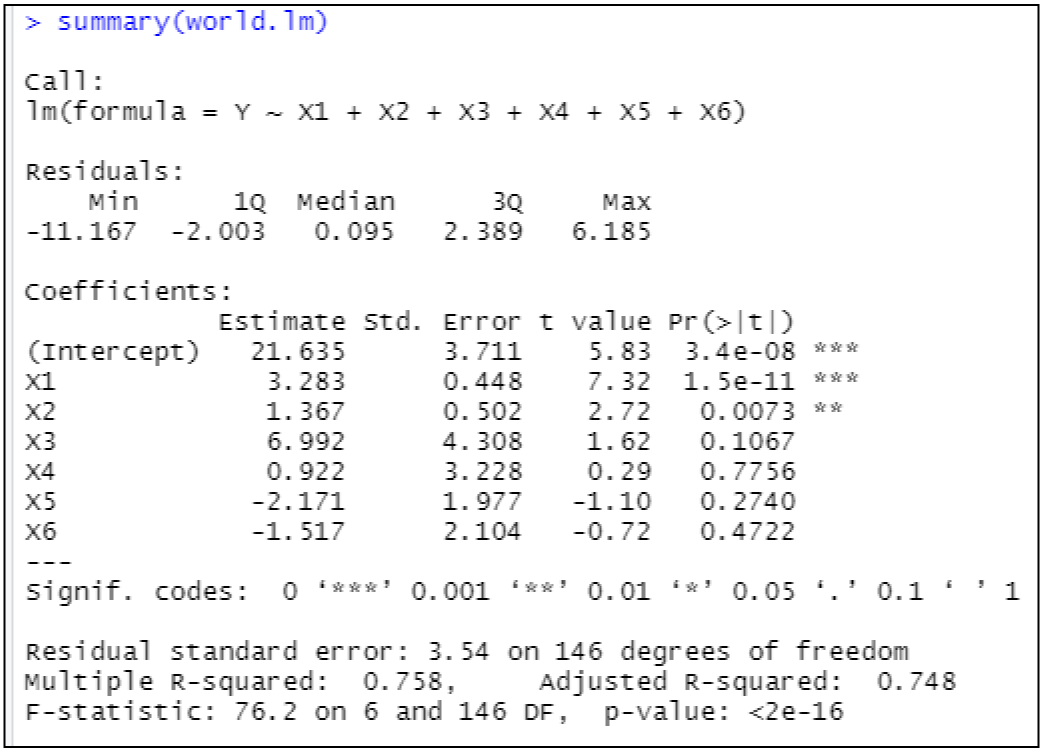


Summary/Anova table full.model

We also used the summary() and anova() functions to get a better understanding of how the data interacts with each other. Our fitted response function is:

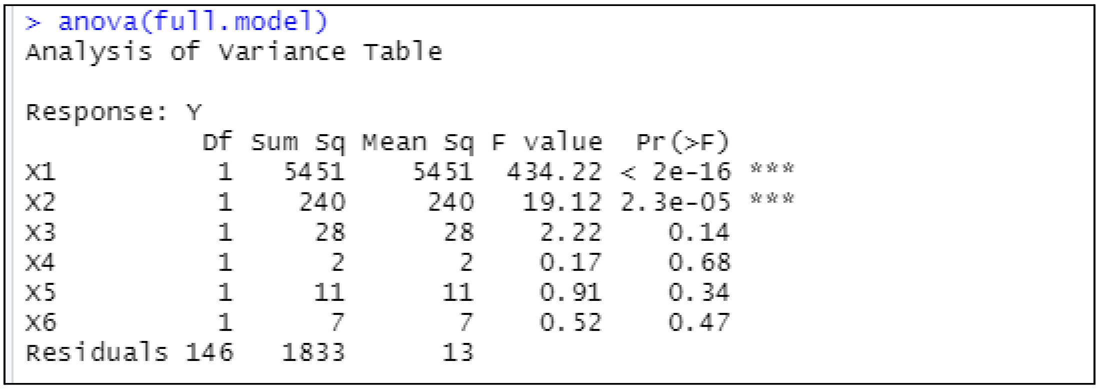
Y = 21.635 + 3.283 (X1) + 1.367 (X2) + 6.992 (X3) + 0.922 (X4) - 2.171 (X5) - 1.517 (X6)

Based on just the assumed p-values (alpha = 0.05), we see that predictors X1 and X2 have the most significance in our data.



```
> summary(world.lm)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -2.003   0.095   2.389   6.185 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   21.635      3.711    5.83  3.4e-08 ***
X1             3.283      0.448    7.32  1.5e-11 ***
X2             1.367      0.502    2.72   0.0073 ** 
X3             6.992      4.308    1.62   0.1067    
X4             0.922      3.228    0.29   0.7756    
X5            -2.171      1.977   -1.10   0.2740    
X6            -1.517      2.104   -0.72   0.4722    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.54 on 146 degrees of freedom
Multiple R-squared:  0.758,     Adjusted R-squared:  0.748 
F-statistic: 76.2 on 6 and 146 DF,  p-value: <2e-16
```

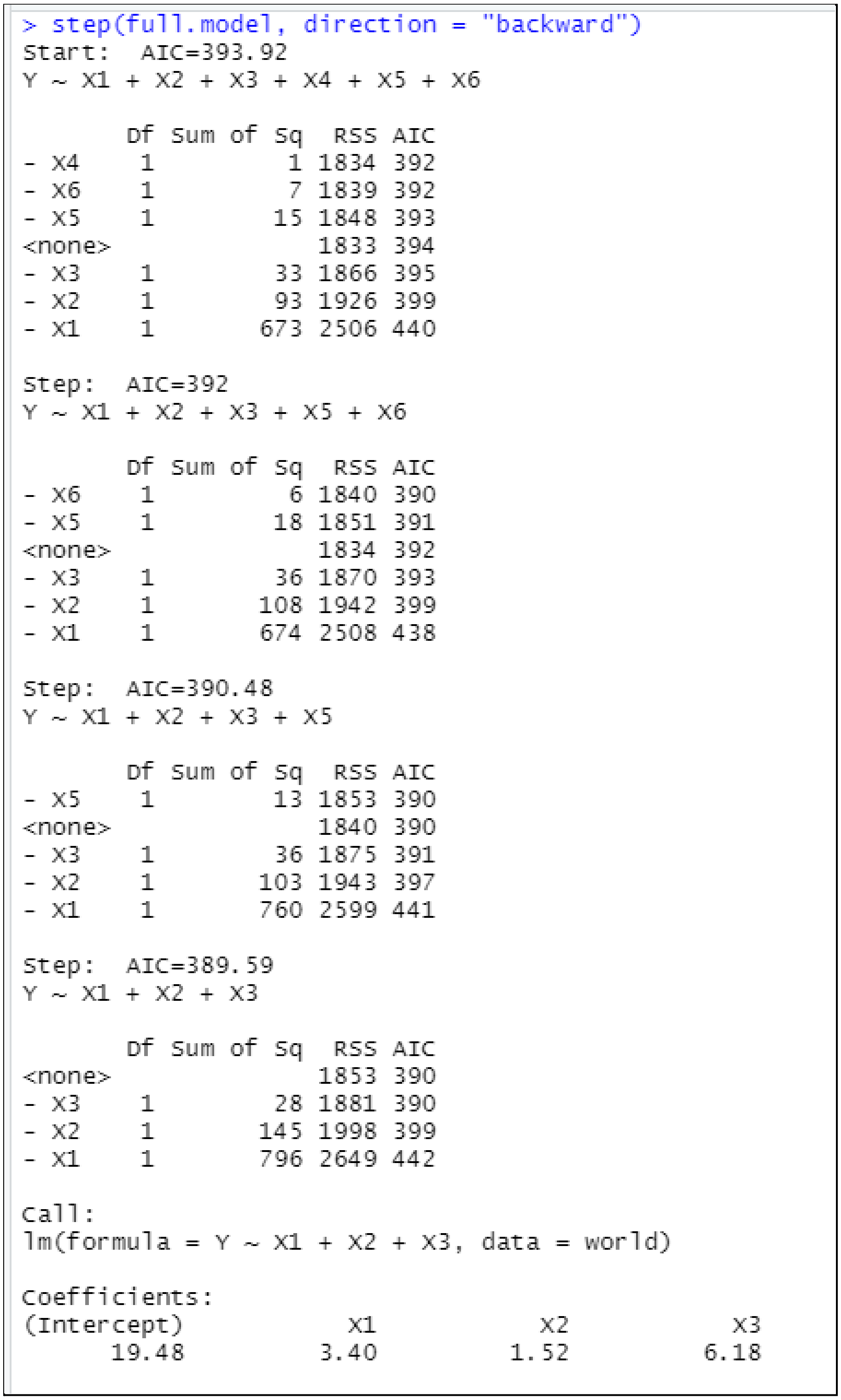
Note: full.model = world.lm just a different name for this part of the project



```
> anova(full.model)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value  Pr(>F)    
X1          1   5451    5451  434.22 < 2e-16 ***
X2          1    240     240   19.12 2.3e-05 ***
X3          1     28      28    2.22    0.14    
X4          1      2       2    0.17    0.68    
X5          1     11      11    0.91    0.34    
X6          1      7       7    0.52    0.47    
Residuals 146   1833      13                    
```

Stepwise

After performing the stepwise regression, function step(), we find that X1, X2, X3 are the most influential predictors in our data. Recall X1= Logged GDP, X2 = Ladder Score, X3 = Social Support



```
> step(full.model, direction = "backward")
Start:  AIC=393.92
Y ~ X1 + X2 + X3 + X4 + X5 + X6

       Df Sum of Sq  RSS AIC
- X4    1         1 1834 392
- X6    1         7 1839 392
- X5    1        15 1848 393
<none>              1833 394
- X3    1        33 1866 395
- X2    1        93 1926 399
- X1    1       673 2506 440

Step:  AIC=392
Y ~ X1 + X2 + X3 + X5 + X6

       Df Sum of Sq  RSS AIC
- X6    1         6 1840 390
- X5    1        18 1851 391
<none>              1834 392
- X3    1        36 1870 393
- X2    1       108 1942 399
- X1    1       674 2508 438

Step:  AIC=390.48
Y ~ X1 + X2 + X3 + X5

       Df Sum of Sq  RSS AIC
- X5    1        13 1853 390
<none>              1840 390
- X3    1        36 1875 391
- X2    1       103 1943 397
- X1    1       760 2599 441

Step:  AIC=389.59
Y ~ X1 + X2 + X3

       Df Sum of Sq  RSS AIC
<none>              1853 390
- X3    1        28 1881 390
- X2    1       145 1998 399
- X1    1       796 2649 442

Call:
lm(formula = Y ~ X1 + X2 + X3, data = world)

Coefficients:
(Intercept)           X1           X2           X3  
      19.48         3.40         1.52         6.18  
```

Hypothesis Testing

For our hypothesis testing, we perform test 1 where we set all the coefficients equal to zero, and test 2 where only the predictors that were eliminated in our stepwise are equal to zero.

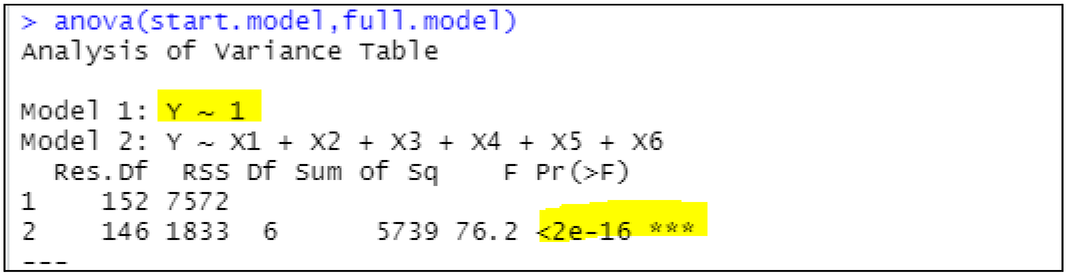
Test 1

H0: B1 = B2 = B3 = B4 = B5 = B6 = 0

Ha: at least one of Bi not equal to zero

full.model <- lm(Y ~ ., data = world)

Reduced.model: start.model <- lm(Y ~ 1, data = world)



```
> anova(start.model,full.model)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1    152 7572
2    146 1833  6      5739 76.2 <2e-16 ***
---
```

n=153, k = 6

F_stat = **76.2**

F_cv = qt(1-alpha, k, n-k-1) = qf(1-0.05, 6, 146) = **2.162**

F_stat = 76.2 > 2.162 = F_cv

P-value= 1-pf(76.2,6,146) = 0

Conclusion: Reject the null hypothesis. Therefore, at least one of B_i not equal to zero. There is sufficient evidence to conclude that life expectancy is significantly related to the 6 predictor variables.

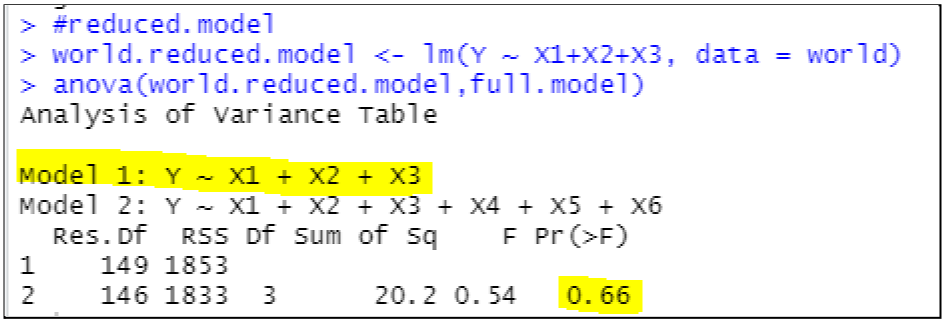
Test 2

H0: B4 = B5 = B6 = 0

Ha: at least one of Bi not equal to zero

full.model <- lm(Y ~ ., data = world)

Reduced.model: world.reduced.model <- lm(Y ~ X1 + X2 + X3, data = world)



```
> #reduced.model
> world.reduced.model <- lm(Y ~ X1+X2+X3, data = world)
> anova(world.reduced.model,full.model)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1    149 1853
2    146 1833  3      20.2 0.54   0.66
```

F_stat = 0.54

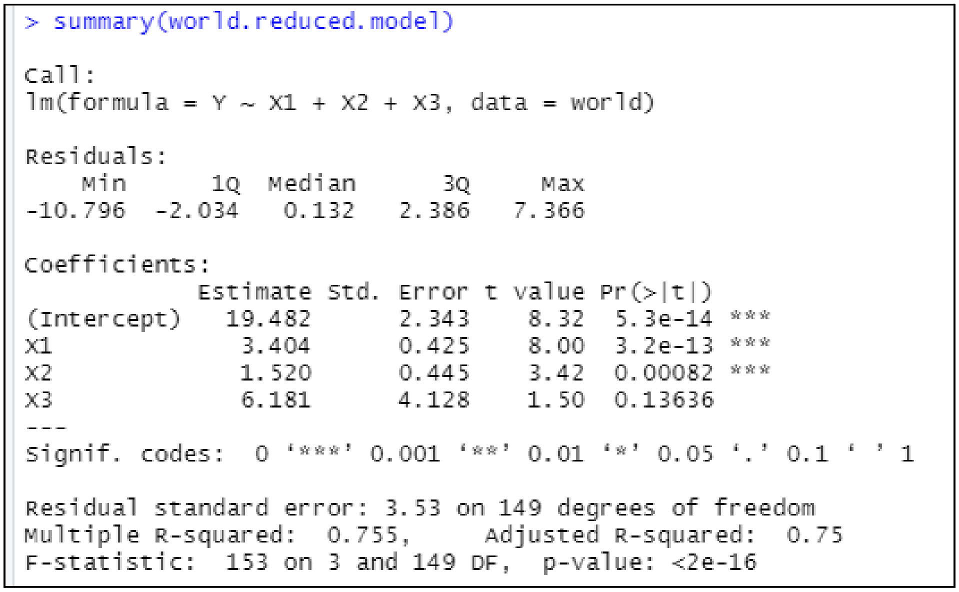
F_cv = qt(1-alpha, k, n-k-1) = qf(1-0.05, 3, 149) = **2.66**

F_stat = 0.54 < 2.162 = F_cv

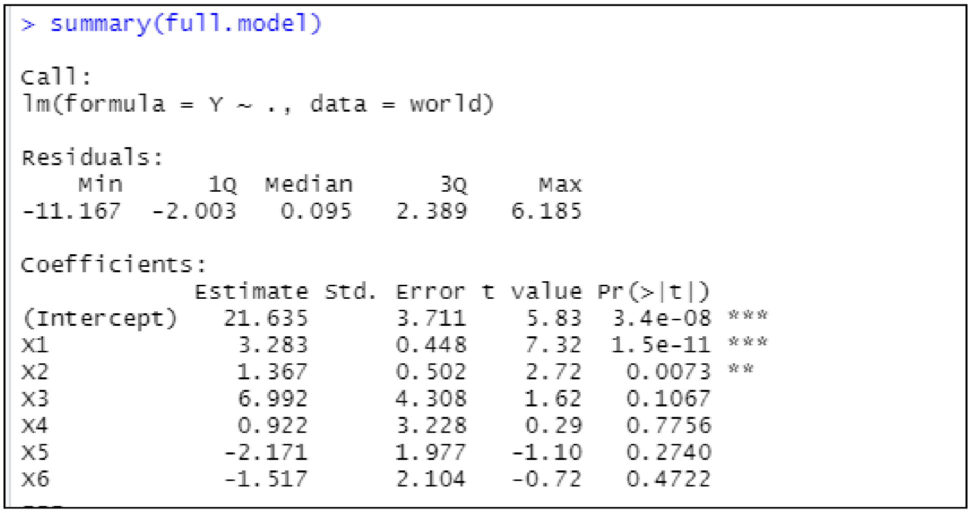
P-value = 1-pf(0.54,3,149) = **0.6556**

Conclusion: We fail to reject the null (the X4, X5, X6 = 0). These 3 predictors are NOT significant in predicting life expectancy.

Finally, we compare the summary() of our reduced model to our full model. We observe that our p-values of the reduced model have increased in significance. Thus further supporting that we have improved our model in predicting the Y response, Life Expectancy.



```
> summary(world.reduced.model)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = world)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.796  -2.034   0.132   2.386   7.366 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   19.482      2.343    8.32  5.3e-14 ***
X1             3.404      0.425    8.00  3.2e-13 ***
X2             1.520      0.445    3.42  0.00082 ***
X3             6.181      4.128    1.50  0.13636    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.53 on 149 degrees of freedom
Multiple R-squared:  0.755,	Adjusted R-squared:  0.75 
F-statistic:  153 on 3 and 149 DF,  p-value: <2e-16
```



```
> summary(full.model)

Call:
lm(formula = Y ~ ., data = world)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -2.003   0.095   2.389   6.185 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   21.635      3.711    5.83  3.4e-08 ***
X1             3.283      0.448    7.32  1.5e-11 ***
X2             1.367      0.502    2.72   0.0073 ** 
X3             6.992      4.308    1.62   0.1067    
X4             0.922      3.228    0.29   0.7756    
X5            -2.171      1.977   -1.10   0.2740    
X6            -1.517      2.104   -0.72   0.4722    
---
```

Conclusions:

As our Conclusion, given our research question, initial expectation - hypothesis in the Introduction and data analysis, we conclude that out of our 6 predictors, **(X1)** Logged of GDP, (**X2**) Ladder score, and (**X3**) Social Support are the most significant predictors to predict **(Y)** Life Expectancy. Compared to our initial expectation - hypothesis when we only use our own subjective opinion without any computation in data analysis, at the end we obtain the same three features as we observed in our class survey.

Along with our progress, we encountered some difficulties such as finding the correct dataset we would like to analyze. In our case, we would like to find the most recent dataset containing actual data in 2020, however, most of the sources on the Internet only contain the World Happiness datasets in the previous years from 2015 up to 2019. To overcome this difficulty, we have visited the main website of the World Happiness Report to find our best fitted dataset. Even though it is stated to be the World Happiness 2020 dataset, it only contains the data up to March 2020 as the most recent time we visited the website, which is not including the data after that date when the pandemic COVID-19 has impacted most of the countries in the world. Thus, we suggest that it would be better for our analysis if we can collect some data after March 2020 which possibly can help us explore more interesting findings such as how COVID-19 would impact people's life.

As we notice throughout our data analysis, 5 out of 6 predictors, (**X2**) Ladder score, (**X3**) Social Support, (**X4**) Freedom to make choices, (**X5**) Perception of Corruption and (**X6**) Generosity are more likely to be collected based on subjective data, which are contained the people's subjective point of view, including feelings, perceptions, and concerns obtained through interviews and surveys. In order to improve the data analysis, a possible procedure we would suggest is that we should have more objective predictors as **(X1)** logged of GDP, which are measurable data obtained through observation, physical examination, country’s economic results, and so on.

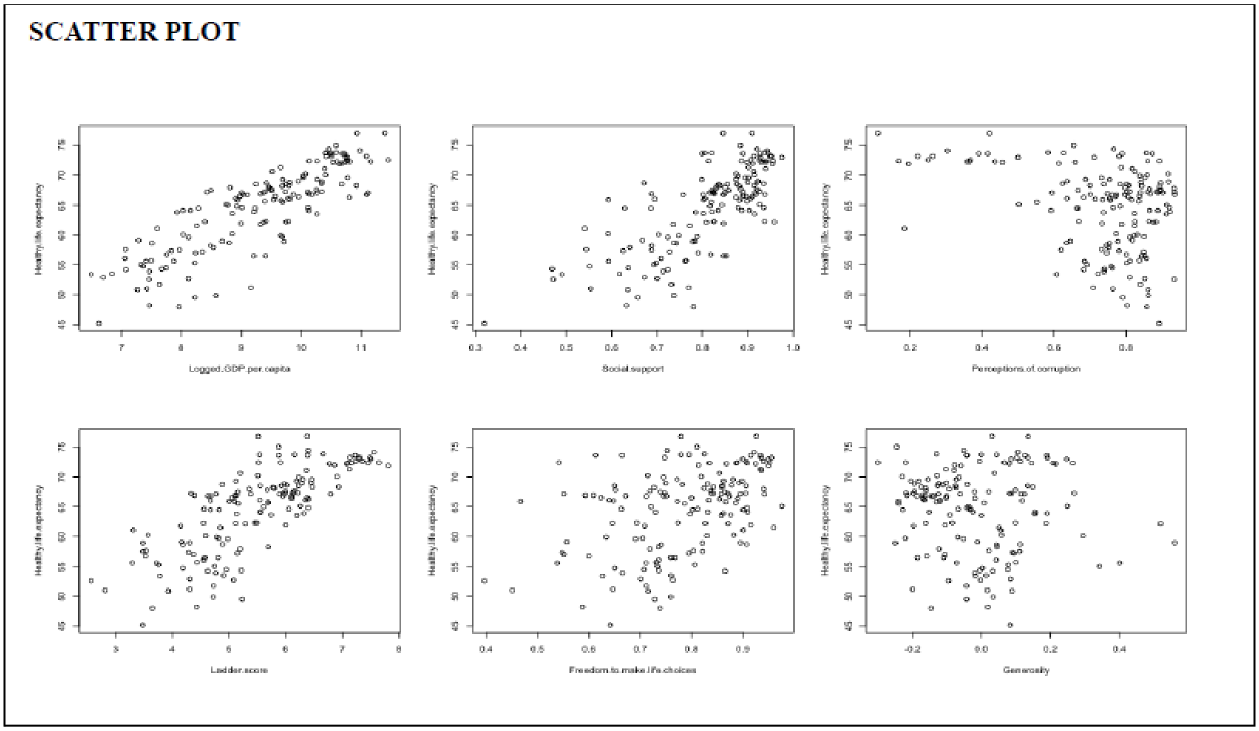
By doing so, it seems more likely to help us have a better result in our data analysis. In addition, as we continue working on our research further in the future, we can potentially learn and apply more approaches and methods to better answer our research question.

References:

- World Happiness Report 2020:
https://www.kaggle.com/mathurinache/world-happiness-report?select=2020.csv
https://worldhappiness.report/ed/2020/#read

Appendix (Optional):

Include all supporting materials, e.g., additional figures/tables, R code.



SCATTER PLOT

R-codes

```
rm(list=ls()) # to remove all objects from a specified environment
cat("\f") # to clean console
library(readr)
data <- read.csv("World_happiness_2020.csv")[,c(-1,-2)]

####################################################################
## Data set up
names(data)
Y = data[,7]  # life expectancy
X1 = data[,5] # Logged GDP
X2 = data[,1] # Ladder score
X3 = data[,6] # Social Support
X4 = data[,8] # Freedom to make choices
X5 = data[,10]# Perception of Corruption
X6 = data[,9] # Generosity
```

```
layout(matrix(c(1:6), 2, 3))

##  correlations/ scatter plot
library(ggplot2)
library(GGally)
ggpairs(data.frame(Y, X1, X2, X3, X4, X5, X6))

pairs(cbind(Y, X1, X2, X3, X4, X5, X6))
plot(X1, Y, xlab = names(data)[5], ylab = names(data)[7])
plot(X2, Y, xlab = names(data)[1], ylab = names(data)[7])
plot(X3, Y, xlab = names(data)[6], ylab = names(data)[7])
plot(X4, Y, xlab = names(data)[8], ylab = names(data)[7])
plot(X5, Y, xlab = names(data)[10], ylab = names(data)[7])
plot(X6, Y, xlab = names(data)[9], ylab = names(data)[7])
par(mfrow=c(1,1))

world =data.frame(Y,X1,X2,X3,X4,X5,X6)
head(world)
world.lm <-lm(Y~X1+X2+X3+X4+X5+X6)
summary(world.lm)
anova(world.lm)

##  Stepwise Regression
full.model  <- lm(Y ~ ., data = world)
start.model <- lm(Y ~ 1, data = world)
step(full.model, direction = "backward")
step(full.model, direction = "both")
step(start.model, direction = "forward", scope = formula(full.model))
step(start.model, direction = "both",    scope = formula(full.model))

##  Hypothesis test
###  Test 1
#   H0: B1 = B2 = B3 = B4 = B5 = B6 = 0
#   Ha: at least one of Bi not equal to zero
#full.model
full.model  <- lm(Y ~ ., data = world)
#reduced.model
start.model <- lm(Y ~ 1, data = world)
anova(start.model,full.model)
###  Test 2
#   H0: B4 = B5 = B6 = 0
#   Ha: at least one of Bi not equal to zero
#full.model
full.model  <- lm(Y ~ ., data = world)
#reduced.model
world.reduced.model <- lm(Y ~ X1+X2+X3, data = world)
anova(world.reduced.model,full.model)
```