Statistics project

Keshav Ganesh
3174597
keshav.ganesh@studbocconi.it

Mattia Barbiere
3151393
mattia.barbiere@studbocconi.it

January 8, 2023

Contents

1 Abstract

How happy is the world? What factors correlate with happiness levels around the world? In this paper we will perform a statistical analysis of the world happiness report, specifically for the years 2021 and 2019.
We obtained our data from *The World Happiness Report* which is a Sustainable Development Solutions Network publication, powered by the Gallup World Poll data. It considers several factors and carries out multiple surveys in each country to generate a happiness score.

By analyzing the years 2021 and 2019, **we will be able to investigate whether the COVID-19 pandemic affected the mean happiness of the world**. First, we will perform model selection to predict the happiness score based on our factors. This will give us an idea of the statistically significant factors. Then, we implement a multi-linear regression to analyze how much variation in the happiness score can be explained by our model. Lastly, we implement a two way-ANOVA to determine if there is a significant difference between the happiness score of each year (our categorical variable being pre-covid and post-covid). We also check if the region of the country has a significant correlation with the happiness level.

1.1 Explaining the Dataset

We are using the dataset submitted on Kaggle under the CC0 Public Domain license. A link to it is available in the bibliography.[1]

We have various factors that can be included in our model. In our table, we have the following attributes:

- **Country, Region**: We divided the countries into 10 regions.
- **Happiness Score**: The rankings of national happiness are based on a Cantril ladder survey[2]. (Basically, a survey asking respondents to rate their happiness score)
- **Logged GDP per Capita**: GDP per Capita is in terms of Purchasing Power Parity (PPP). We use the natural log of GDP per Capita, as this form fits the data significantly better than GDP per Capita.
- **Social Support**: The national average of the binary responses (0=no, 1=yes) to the 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?"
- **Healthy Life Expectancy**: The time series for healthy life expectancy at birth is based on data from the World Health Organization (WHO).
- **Freedom to make life choices**: Freedom to make life choices is the national average of binary responses (0=no, 1=yes) to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
- **Generosity**: Generosity is the residual of regressing the national average of GWP responses to the donation question "Have you donated money to a charity in the past month?" on log GDP per Capita.
- **Perception of corruption**: Perceptions of corruption are the average of binary answers to two questions: "Is corruption widespread throughout the government in this country or not?" and "Is corruption widespread within businesses in this country or not?"

This data was collected by Gallup, Inc in the "Gallup World Poll" and outlined in the world happiness report published every year. It also includes more details on the definitions and the survey methods.[3]

We only included the countries that were surveyed in both 2019 and 2021 to simplify the statistical analysis. In the end, 144 countries were used in the investigation. Unfortunately, many countries had some missing data (especially in 2019) and this caused problems with the model selection code. Therefore, we had to omit those countries at least for the model selection procedure. In the end, we included 126 countries in the 2019 model selection while all 144 countries were included in the 2021 model selection.

2 Model Selection

We set out to create a multilinear model with happiness as the dependent variable. We aim to create the smallest model that contains all the statistically relevant covariates. To do so, we implement some model selection methods. For the 2019 data, we used the BIC and step-up methods and for the 2021 data, we used the BIC and step-down methods. In both cases, the final model coincided. It should be noted that we did try step-down for 2019 and step-up for 2021 but they gave us different results from the other two methods therefore, we will not include those cases in our report. In order to build a regression model we will assume that all the residuals are normally distributed with zero mean and constant variance. In section 2.4 we will discuss whether these assumptions are met with our data. Please note that we choose to multiply Social Support and Healthy Life Expectancy to get a more constant variance.

Additionally, we attempted to implement model selection with the region parameter included. It gave a slightly larger value of R^2, but it adds a lot of confusion to the output. So, we decided to preclude it from the investigation. Any effects brought about by the region are noticed when we implement ANOVA in section 4.

2.1 Step-Down for 2021 data

In the step-down method, we started off with the biggest possible model (modelMax21) and with some iteration, we discarded the covariate with the least amount of statistical significance in our model. The step() function[4] uses the AIC value as the criteria for elimination. In appendix A you can find the console log of the function. We also added the p-values of each covariate to confirm that the order of elimination was correct. This is the function that did all the heavy lifting in the step-down, and step-up, methods.

```
###Step-Down for 2021###
step(modelMax21, direction="backward", test="F")
```

Using the step-down method we conclude that the best linear model is:

- Logged GDP per Capita
- Social Support
- Social Support * Healthy Life Expectancy
- Freedom to make life choices
- Healthy Life Expectancy

2.2 Step-Up for 2019 data

For the 2019 data, we used a step-up method, which starts with an empty model (modelStart19) and gradually adds covariates by choosing the ones with the lowest p-value or, in our case, the AIC value; nevertheless, the result is identical. We used the step() function[5] again but we had to specify the largest acceptable model (modelMax19).

```
###Step-Up for 2019###
step(modelStart19, direction="forward", scope=formula(modelMax19), test="F")
```

Using the step-up method we get the following model:

- Social Support * Healthy Life Expectancy
- Freedom to make life choices
- Perceptions of corruption

2.3 BIC

Our final method of model selection is the Bayesian Information Criterion(BIC). Using the BIC we could compare various models and ultimately find the one that minimized the BIC. To implement this we looped through all the possible models. This method is not very efficient when you have large amounts of covariates, but in our case, we only had 7. We refer the reader to Appendix D for the BIC code.

We implemented the BIC method both for 2019 and 2021:

2021:
- Logged GDP per Capita
- Social Support
- Social Support * Healthy Life Expectancy
- Freedom to make life choices
- Healthy Life Expectancy

2019:
- Social Support * Healthy Life Expectancy
- Freedom to make life choices
- Perceptions of corruption

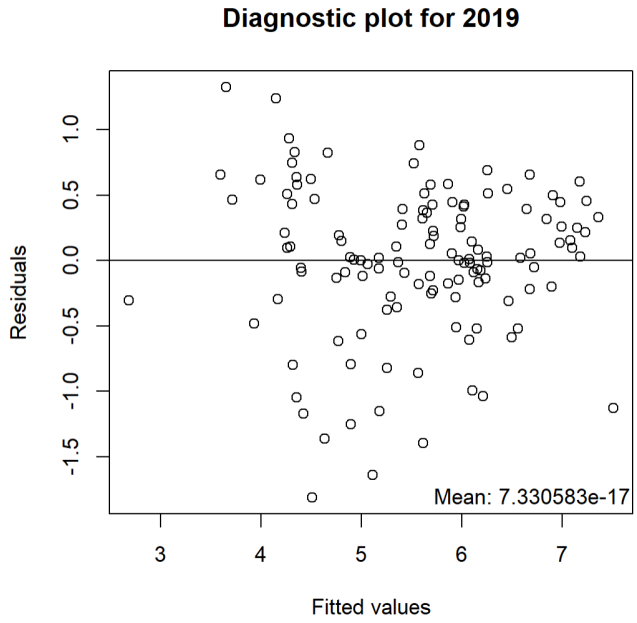
We have found a possible regression model for each year, but before the multilinear regression let's check if our initial assumptions are satisfied.

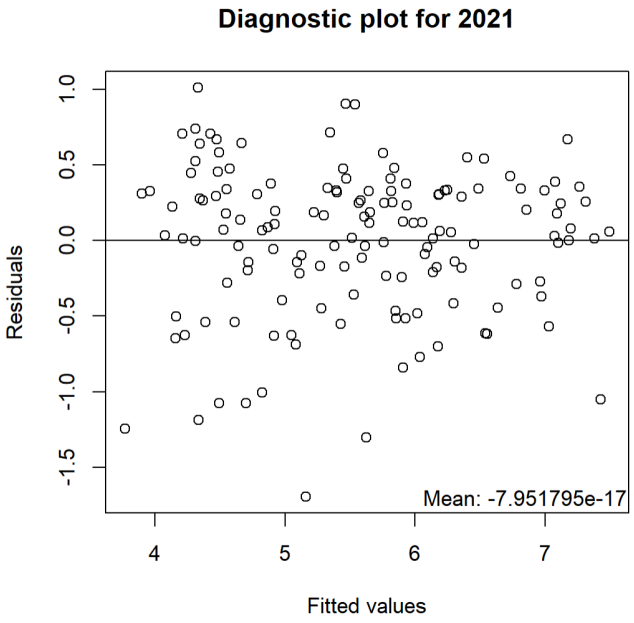
2.4 Regression Diagnostics

We started our model selection with three assumptions; the residuals are normally distributed with mean zero and constant variance.

2.4.1 Mean and Variance analysis

As you saw in the previous section, we added a covariate which was the product of Social Support and Healthy Life expectancy. In fact, this covariate appears in both models after the model selection. Consequently, we get a more constant variance when this independent variable is used.
Below you can find the plot of the dependent variables (Fitted) against the residuals:

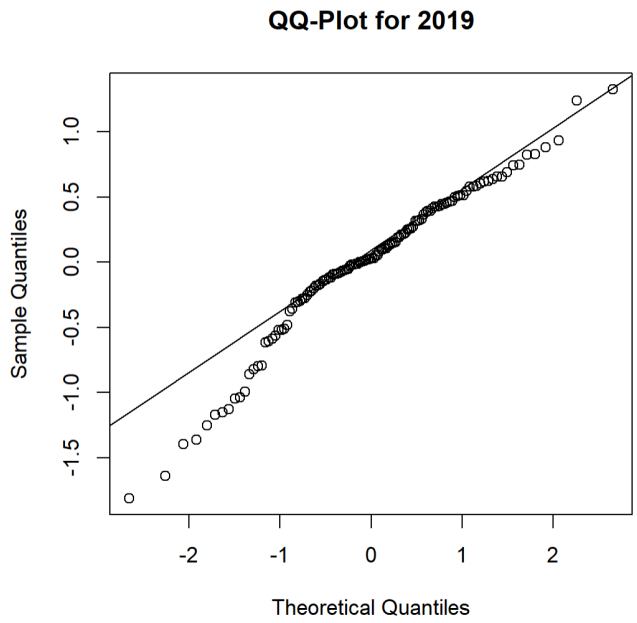


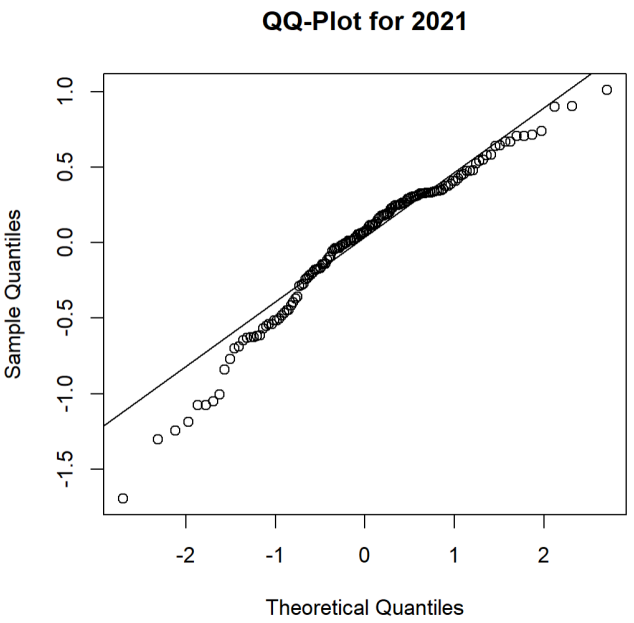


From a visual inspection of these plots, we can see that the zero mean assumption is reasonable for both years. When it comes to constant variance in 2021, we see that overall, the assumption seems to be satisfied with some small deviations around the fitted values of 5 and 6. So the 2021 data is in line with the constant variance assumption. On the other hand, in 2019, we notice a very evident deviation from the constant variance assumption. We can discern somewhat of a parabola symmetric around 5. This is a clear indication that for 2019, a non-linear model would probably be a better fit. So even though we will assume constant variance in our multilinear analysis for the 2019 data, this assumption is not entirely true. This should be taken into account when we present conclusions taken from the regression model.

2.4.2 Normality analysis

When it comes to the normality assumption we can make QQ-plots for a visual inspection:





From the QQ-plots, we deduce that in 2019 we have a substantial amount of points that deviate from the line, especially in the lower part of the plot.
In 2021, on the other hand, we have only a few points that stray away from the line, but the evidence for normality is still much stronger than in 2019.
When it comes to the Shapiro-Wilk test, however, for both years, we get a very small p-value (of the order of 10^{-4}) which is evidence against normality.
In conclusion, the normality assumption is not completely absurd with our data and in fact, for 2021 it seems to be quite reasonable.
For 2019 however, the assumption deviates somewhat from the real data, so any conclusion taken from the multilinear regression for the 2019 data should be taken with a grain of salt.

3 Multilinear Regression

With the model selection and regression diagnostic completed, we can officially start with our multilinear regression with the models found in the previous section. Utilizing the built-in R function for linear regression we are ready to analyze the output.

3.1 Analysis of Multilinear Regression

You can find the results of our regression in Appendix B. We get these fitted regression models:

2019

$$\widehat{Happiness} = \begin{bmatrix} 1.310 \\ 0.065 \\ 1.717 \\ -0.800 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ \text{LifeExpectancyTimesSocialSupport} \\ \text{FreedomToMakeLifeChoices} \\ \text{PerceptionsOfCorruption} \end{bmatrix}$$

2021

$$\widehat{Happiness} = \begin{bmatrix} 15.249 \\ 0.187 \\ -20.508 \\ 2.456 \\ -0.255 \\ 0.371 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ \text{LoggedGDPperCapita} \\ \text{SocialSupport} \\ \text{FreedomToMakeLifeChoices} \\ \text{HealthyLifeExpectancy} \\ \text{LifeExpectancyTimesSocialSupport} \end{bmatrix}$$

For both years all the covariates have statistical evidence (with $\alpha = 0.05$) that they should be included in the model. This is of course a direct consequence of our model selection procedure.
Secondly, our eyes are immediately drawn to the extremely low p-value for the F-test. This tells us that both models are better than an empty model, which makes sense since the variables in our models are bound to have some correlation with happiness. For instance, intuitively, the *Freedom To Make Life Choices* should have some correlation with happiness.
Moving to the R-squared values, here we notice that 2021 and 2019 have an R-squared value of approximately 0.79 and 0.74 respectively. This tells us that about 79% (resp. 74%) of the variations in happiness are explained by our independent variables. This seems to be reasonably correct as many more factors influence people's happiness such as having a stable job or access to food and water. However, our few covariates still manage to explain about 3/4 of the variations which is already a substantial amount. Adjusting the R-squared only changes the value by about 1% for both years.

4 ANOVA

We have finally arrived at the core of our research question. We are now going to see if COVID-19 influenced people's happiness. To do this we do an ANOVA using the year as a categorical variable. Some people may argue that COVID-19 might have still been present in some counties in 2021 but as of the time of writing this is the most up-to-date data set we have. Notably, for this section, we choose to use our entire data set as every country has a happiness score. Therefore, there is no need to eliminate rows with missing data. We also combined the data and added a column for the year to execute the in-built function. In addition, we included the Region to see if it has a significant effect and in fact, this is exactly what we see.

Our results:

```
                            Df Sum Sq Mean Sq F value Pr(>F)
factor(year)                 1   0.02   0.017   0.037  0.847
factor(Region)               9 214.61  23.846  51.901 <2e-16 ***
factor(year):factor(Region)  9   0.26   0.029   0.063  1.000
Residuals                  268 123.13   0.459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We clearly see that the year and the interaction effects have a relatively high p-value and so are not statistically significant(again with $\alpha = 0.05$). We will dive deeper into this in the next section. On the flip side, the Region variable has an extremely low p-value, which is in line with the unfortunate fact that the region where you live has a dramatic effect on your social and economic situation which in turn should positively correlate with happiness.

5 Conclusions and findings

Happiness is a very complicated topic to analyze from a statistical point of view as there are an absurd amount of factors affecting it. In reality, said factors aren't necessarily common across all humans. Culture and religion might influence what people need to be happy, so it becomes very complicated to represent all these factors in a single data set. For these reasons, it might be very hard to find a single model that works for every country. Hope is not lost, however, as social and economic factors can already give us a good understanding. Even though people might have different wants we are all humans and we all have the same needs in life. So, we can still attempt to create a model for happiness using our data, and that is exactly what we did.

We abstracted from all these considerations and worked with our very limited data set. Some might argue that this oversimplifies the analysis but, as we have already said, a perfect model is borderline impossible. Let's take a closer look at our model selection results. We see that the best model does change over the year. This does not mean that those factors are the only factors that correlate with people's happiness; they are just the best combination of factors to model the data we have. In 2019, for example, the model selection does not include Logged GDP per Capita, but more money usually makes people happier(at least up to a certain point). However, for our data evidently, its influence was not large enough to be considered.

Unfortunately, our study is not a good method to understand what makes people happy because correlation does not imply causation. Rather it is a model that predicts the data the best. We cannot make statistical conclusions about causation but many of these factors do have a causal effect on people's happiness and in fact, this is the reason why they were included in the data set to begin with.

We finally arrive at our main question. Did COVID-19 have a statistically significant effect on people's happiness? A lot of people might say yes to this question, however, we are looking for a more rigorous answer. In fact, our ANOVA tells us that the year does not have a statistically significant effect. So our statistical investigation comes out negative. The main reason we can think of is that people may have become less happy during the various restrictions in their respective countries, but once the pandemic came to an end, the happiness score returned more or less to the pre-pandemic levels. There is a slight decrease in the happiness score mean between 2019 and 2021 with a difference of the order 10^{-2}, a very small difference. A graph outlining this difference can be seen in Appendix C. Then again, COVID-19 hit different countries in different ways and at different times, and we don't think yearly data is precise enough to see the fluctuations in happiness due to COVID-19. So happiness remains a very complicated topic and it is still very hard to understand and predict the world's happiness, but as long as people are happy does it really matter why?

A Model Selection

A.1 Stepwise Regression

A.1.1 Step-up method 2019

```
Start:  AIC=31.87
HappinessScore ~ 1

                                    Df Sum of Sq     RSS      AIC  F value    Pr(>F)
+ LifeExpectancyTimesSocialSupport  1   110.189  49.514 -113.686 275.9489 < 2.2e-16 **
*
+ HealthyLifeExpectancy             1    95.020  64.683  -80.014 182.1569 < 2.2e-16 **
*
+ LoggedGDPperCapita                1    91.864  67.840  -74.011 167.9130 < 2.2e-16 **
*
+ SocialSupport                     1    89.919  69.785  -70.448 159.7748 < 2.2e-16 **
*
+ FreedomToMakeLifeChoices          1    51.710 107.994  -15.430  59.3740 3.610e-12 **
*
+ PerceptionsOfCorruption           1    31.144 128.560    6.534  30.0393 2.265e-07 **
*
<none>                                          159.704   31.867
+ Generosity                        1     0.001 159.703   33.866   0.0006      0.98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-113.69
HappinessScore ~ LifeExpectancyTimesSocialSupport

                           Df Sum of Sq    RSS     AIC F value    Pr(>F)
+ FreedomToMakeLifeChoices  1    6.3882 43.126 -129.09 18.2196 3.895e-05 ***
+ PerceptionsOfCorruption   1    5.1251 44.389 -125.45 14.2012 0.0002536 ***
+ SocialSupport             1    1.6472 47.867 -115.95  4.2327 0.0417659 *
+ Generosity                1    1.5904 47.924 -115.80  4.0819 0.0455166 *
<none>                                  49.514 -113.69
+ HealthyLifeExpectancy     1    0.7399 48.775 -113.58  1.8659 0.1744427
+ LoggedGDPperCapita        1    0.4141 49.100 -112.74  1.0375 0.3104086
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-129.09
HappinessScore ~ LifeExpectancyTimesSocialSupport + FreedomToMakeLifeChoices

                           Df Sum of Sq    RSS     AIC F value  Pr(>F)
+ PerceptionsOfCorruption  1   2.25015 40.876 -133.84  6.7159 0.01072 *
+ SocialSupport            1   1.20228 41.924 -130.65  3.4987 0.06381 .
<none>                                 43.126 -129.09
+ LoggedGDPperCapita       1   0.59921 42.527 -128.85  1.7190 0.19229
+ Generosity               1   0.41974 42.707 -128.32  1.1991 0.27567
+ HealthyLifeExpectancy    1   0.39473 42.732 -128.25  1.1270 0.29052
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-133.84
HappinessScore ~ LifeExpectancyTimesSocialSupport + FreedomToMakeLifeChoices +
    PerceptionsOfCorruption

                          Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>                                 40.876 -133.84
+ SocialSupport           1   0.52474 40.351 -133.47  1.5735 0.2121
+ LoggedGDPperCapita      1   0.27898 40.597 -132.71  0.8315 0.3637
+ HealthyLifeExpectancy   1   0.15959 40.717 -132.34  0.4743 0.4924
+ Generosity              1   0.11135 40.765 -132.19  0.3305 0.5664

Call:
lm(formula = HappinessScore ~ LifeExpectancyTimesSocialSupport +
    FreedomToMakeLifeChoices + PerceptionsOfCorruption, data = data19)

Coefficients:
                    (Intercept)  LifeExpectancyTimesSocialSupport
                         1.3104                            0.0651
        FreedomToMakeLifeChoices           PerceptionsOfCorruption
                         1.7168                           -0.8002
```

A.1.2 Step-down method 2021

```
Start:  AIC=-194.03
HappinessScore ~ (`Country name` + Region + year + LoggedGDPperCapita +
    SocialSupport + HealthyLifeExpectancy + FreedomToMakeLifeChoices +
    Generosity + PerceptionsOfCorruption + LifeExpectancyTimesSocialSupport) -
    `Country name` - Region


Step:  AIC=-194.03
HappinessScore ~ LoggedGDPperCapita + SocialSupport + HealthyLifeExpectancy +
    FreedomToMakeLifeChoices + Generosity + PerceptionsOfCorruption +
    LifeExpectancyTimesSocialSupport

                                  Df Sum of Sq    RSS     AIC F value    Pr(>F)
- PerceptionsOfCorruption          1    0.0000 33.490 -196.03  0.0000   0.99653
- Generosity                       1    0.2175 33.707 -195.10  0.8833   0.34898
<none>                                        33.490 -194.03
- LoggedGDPperCapita               1    1.4735 34.963 -189.83  5.9838   0.01572 *
- HealthyLifeExpectancy            1    4.9326 38.423 -176.25 20.0311 1.597e-05 ***
- SocialSupport                    1    5.1053 38.595 -175.60 20.7324 1.161e-05 ***
- FreedomToMakeLifeChoices         1    6.1472 39.637 -171.77 24.9633 1.764e-06 ***
- LifeExpectancyTimesSocialSupport  1    6.6243 40.114 -170.04 26.9007 7.597e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-196.03
HappinessScore ~ LoggedGDPperCapita + SocialSupport + HealthyLifeExpectancy +
    FreedomToMakeLifeChoices + Generosity + LifeExpectancyTimesSocialSupport

                                  Df Sum of Sq    RSS     AIC F value    Pr(>F)
- Generosity                       1    0.2212 33.711 -197.09  0.9050   0.34311
<none>                                        33.490 -196.03
- LoggedGDPperCapita               1    1.4861 34.976 -191.78  6.0793   0.01491 *
- HealthyLifeExpectancy            1    5.5766 39.066 -175.85 22.8125 4.534e-06 ***
- SocialSupport                    1    6.0554 39.545 -174.10 24.7712 1.906e-06 ***
- FreedomToMakeLifeChoices         1    6.7562 40.246 -171.57 27.6383 5.482e-07 ***
- LifeExpectancyTimesSocialSupport  1    7.7029 41.193 -168.22 31.5108 1.061e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=-197.09
HappinessScore ~ LoggedGDPperCapita + SocialSupport + HealthyLifeExpectancy +
    FreedomToMakeLifeChoices + LifeExpectancyTimesSocialSupport

                                  Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                                        33.711 -197.09
- LoggedGDPperCapita               1    1.3391 35.050 -193.48  5.4818   0.02065 *
- HealthyLifeExpectancy            1    6.1542 39.865 -174.94 25.1928 1.573e-06 ***
- SocialSupport                    1    6.5955 40.307 -173.35 26.9993 7.162e-07 ***
- FreedomToMakeLifeChoices         1    8.3351 42.046 -167.27 34.1205 3.549e-08 ***
- LifeExpectancyTimesSocialSupport  1    8.3734 42.085 -167.14 34.2774 3.327e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = HappinessScore ~ LoggedGDPperCapita + SocialSupport +
    HealthyLifeExpectancy + FreedomToMakeLifeChoices + LifeExpectancyTimesSocialSuppor
t,
    data = data21)

Coefficients:
                      (Intercept)                LoggedGDPperCapita
                          15.2486                            0.1871
                   SocialSupport             HealthyLifeExpectancy
                          -20.5084                           -0.2550
         FreedomToMakeLifeChoices  LifeExpectancyTimesSocialSupport
                           2.4558                            0.3713
```

B Multilinear Regression Results

2019

```
Call:
lm(formula = HappinessScore ~ LifeExpectancyTimesSocialSupport +
    FreedomToMakeLifeChoices + PerceptionsOfCorruption, data = data19)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81360 -0.22471  0.02753  0.40847  1.32810

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.310420   0.541792   2.419  0.01705 *
LifeExpectancyTimesSocialSupport 0.065105   0.004877  13.350  < 2e-16 ***
FreedomToMakeLifeChoices         1.716781   0.530168   3.238  0.00155 **
PerceptionsOfCorruption         -0.800152   0.308761  -2.591  0.01072 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5788 on 122 degrees of freedom
Multiple R-squared:  0.7441,    Adjusted R-squared:  0.7378
F-statistic: 118.2 on 3 and 122 DF,  p-value: < 2.2e-16
```

2021

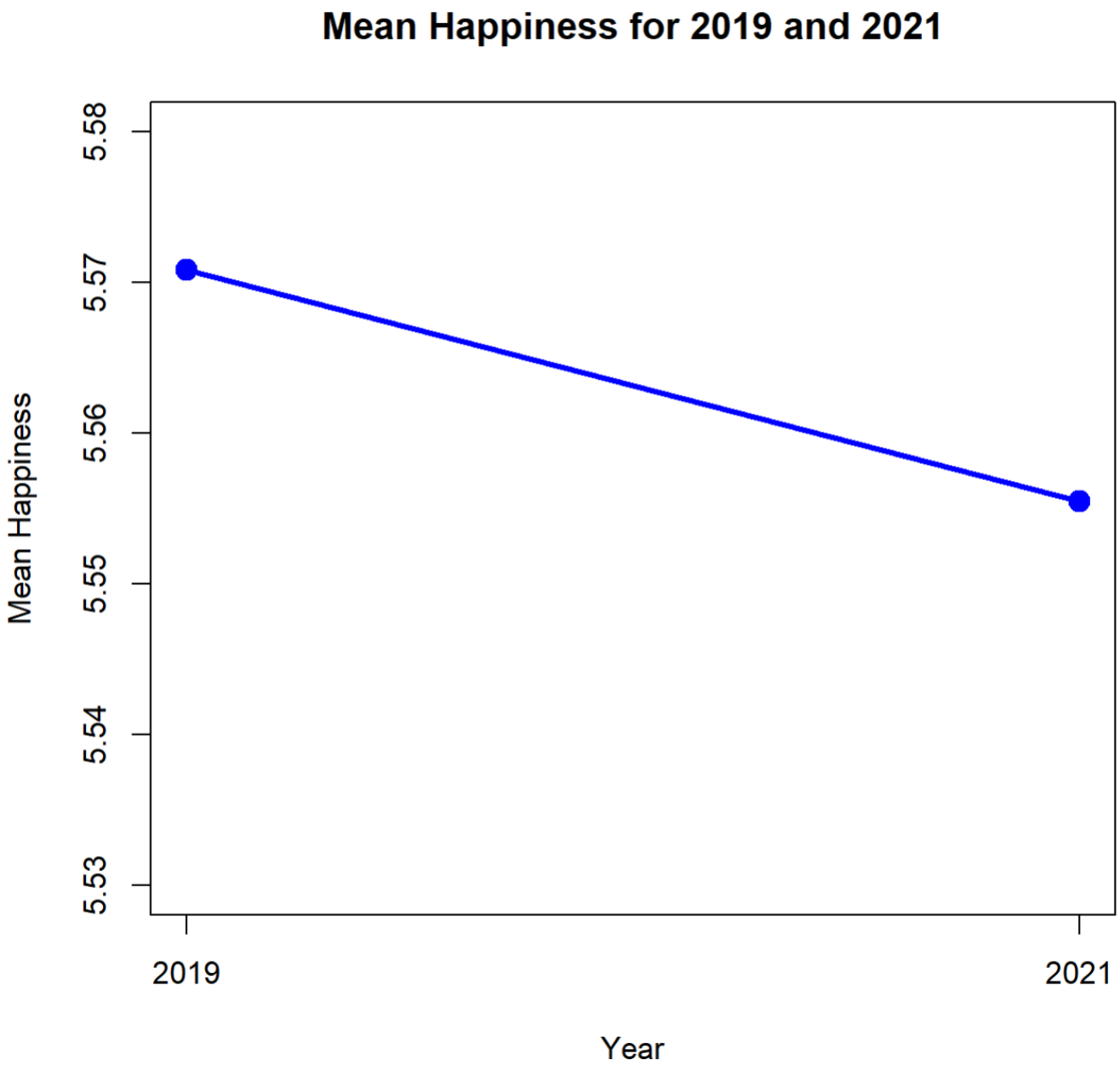
```
Call:
lm(formula = HappinessScore ~ LoggedGDPperCapita + SocialSupport +
    FreedomToMakeLifeChoices + HealthyLifeExpectancy + LifeExpectancyTimesSocialSupport,
    data = data21)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69483 -0.25171  0.07533  0.32708  1.01061

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      15.24863    3.18510   4.787 4.29e-06 ***
LoggedGDPperCapita                0.18712    0.07992   2.341   0.0206 *
SocialSupport                   -20.50841    3.94690  -5.196 7.16e-07 ***
FreedomToMakeLifeChoices          2.45578    0.42042   5.841 3.55e-08 ***
HealthyLifeExpectancy            -0.25504    0.05081  -5.019 1.57e-06 ***
LifeExpectancyTimesSocialSupport  0.37132    0.06342   5.855 3.33e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4943 on 138 degrees of freedom
Multiple R-squared:  0.7909,    Adjusted R-squared:  0.7833
F-statistic: 104.4 on 5 and 138 DF,  p-value: < 2.2e-16
```

C Mean Happiness



D BIC Code

```
###BIC###
data19 <- Reports_final_2019_Ganesh_Barbiere
data21 <- Reports_final_2021_Ganesh_Barbiere
linear_model_generator <- function(x){
  attribute_list = list()
  if (x[1] == 1){
    attribute_list <- append(attribute_list, "LoggedGDPperCapita")
  }
  if (x[2] == 1){
    attribute_list <- append(attribute_list, "FreedomToMakeLifeChoices")
  }
  if (x[3] == 1){
    attribute_list <- append(attribute_list, "Generosity")
  }
  if (x[4] == 1){
    attribute_list <- append(attribute_list, "PerceptionsOfCorruption")
  }
  if (x[5] == 1){
    attribute_list <- append(attribute_list, "LifeExpectancyTimesSocialSupport")
  }
  if (x[6] == 1){
    attribute_list <- append(attribute_list, "HealthyLifeExpectancy")
  }
  if (x[7] == 1){
    attribute_list <- append(attribute_list, "SocialSupport")
  }
  return(attribute_list)
}

n <- 7
l <- rep(list(0:1), n)

df = expand.grid(l)

##BIC2019##

empty_2019 <- lm(`HappinessScore` ~ 1, data=data19)

BIC_list_2019 <- list()

BIC_list_2019 <- list(1:128)
BIC_list_2019[1] <- BIC(empty_2019)

for (j in 2:128){
  frmla <- as.formula(paste("HappinessScore", paste(linear_model_generator(df[j,]), sep = "",
                                                    collapse = " + "), sep = " ~ "))
  my_model = lm(formula = frmla, data = data19)
  BIC_list_2019[j] = BIC(my_model)
}

v = unlist(BIC_list_2019)
index = which(v == min(v))
frmla19 <- as.formula(paste("HappinessScore", paste(linear_model_generator(df[index,]), sep = "",
                                                    collapse = " + "), sep = " ~ "))

optimal_BICmodel_2019 = lm(formula = frmla19, data = data19)

summary(optimal_BICmodel_2019)
```

```
##BIC2021##

empty_2021 <- lm(`HappinessScore` ~ 1, data=data21)

BIC_list_2021 <- list()

BIC_list_2021 <- list(1:128)
BIC_list_2021[1] <- BIC(empty_2021)

for (j in 2:128){
  frmla <- as.formula(paste("HappinessScore", paste(linear_model_generator(df[j,]), sep = "",
                                                    collapse = " + "), sep = " ~ "))
  my_model = lm(formula = frmla, data = data21)
  BIC_list_2021[j] = BIC(my_model)
}

w = unlist(BIC_list_2021)
index = which(w == min(w))
frmla21 <- as.formula(paste("HappinessScore", paste(linear_model_generator(df[index,]), sep = "",
                                                    collapse = " + "), sep = " ~ "))

optimal_BICmodel_2021 = lm(formula = frmla21, data = data21)
summary(optimal_BICmodel_2021)
```

E Bibliography

References

[1] Ajaypal Singh.
Source: *World Happiness Report 2021*

[2] Gallup, Inc.
Source: *Understanding How Gallup Uses the Cantril Scale*

[3] Gallup, Inc.
Source: *More details on the definitions*

[4] Dragonfly Statistics
Source: *Backward Elimination*

[5] Dragonfly Statistics
Source: *Forward Selection*