Predicting Happiness Score with Multiple Linear Regression Model

Kuan-Yi Chen, Naomi Chou

Introduction

Background Information

Our data is drawn from the "World Happiness Report", an annual publication that ranks the happiness levels of countries based on six main factors – Gross Domestic Product (GDP) per capita, healthy life expectancy, social support, freedom to make life choices, generosity, and perceptions of corruption. This report specifically analyzes data from 2019, encompassing 156 countries.

Research Question

In this dataset, our goal is to identify which of the six predictor variables – be it GDP per capita, freedom to make life choices, social support, or another factor – has the most significant influence on the happiness score.

Method

In the scope of our research, we utilized hypothesis testing as a means of discerning the optimal fit of the data, specifically whether a reduced or full model is more suitable. Following this, we employed methodologies like the inverse response plot and the application of the Box-Cox transformation to achieve the necessary model transformations. Then, the process of variable selection was determined through a comprehensive array of techniques, including the comparative analysis of all possible subsets and stepwise regression.

Overview

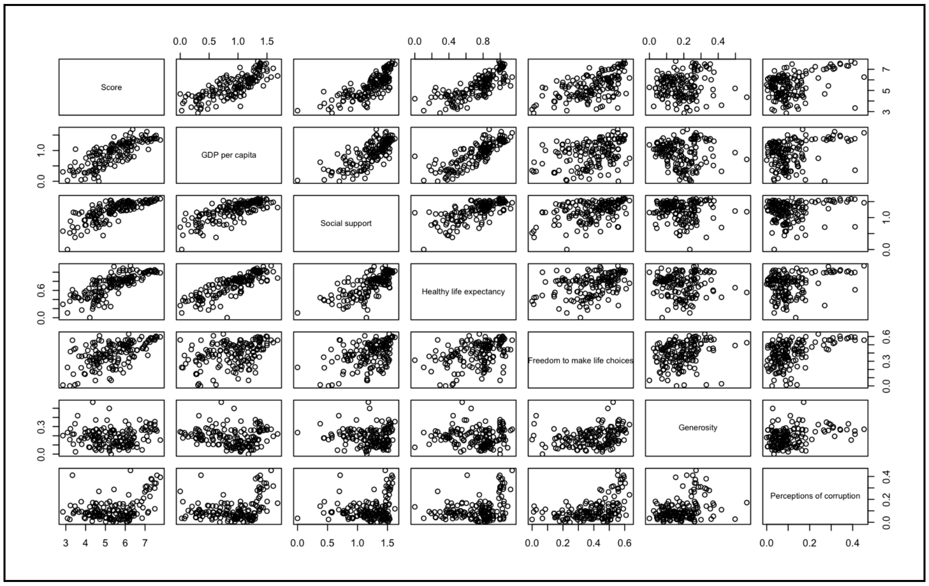
This paper begins with an introduction that outlines the research objectives and different methodologies utilized for data analysis. In the data description and interpretation section, the full model analysis highlights the weaknesses of the model, allowing us to determine the necessary transformations. A detailed analysis of the summary statistics and diagnostic plots and tools is presented. Through various model validation methods, we choose the optimal model with four predictors – GDP per capita, social support, healthy life expectancy, and freedom to make life choices. This report ends with a discussion of the limitations and potential improvements of our data and analysis.

Data Description and Interpretation

Defining the variables

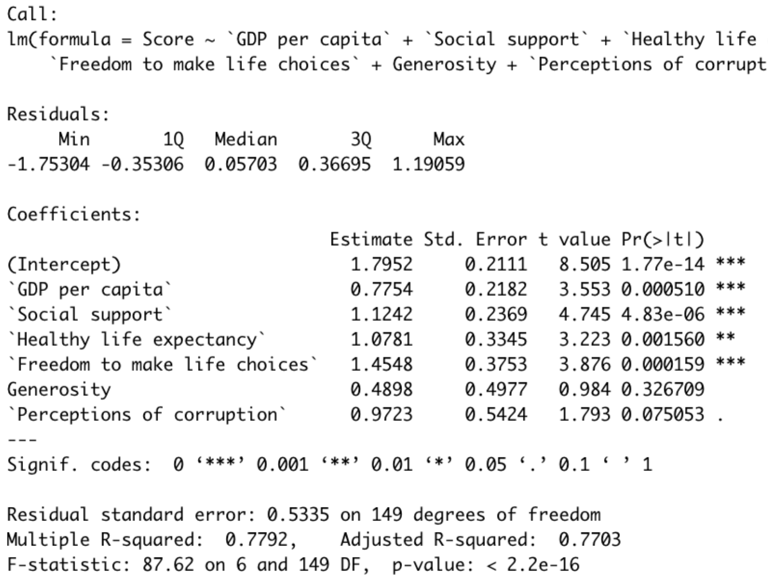
Y = Score
X_1 = GDP per capita
X_2 = Social support
X_3 = Healthy life expectancy
X_4 = Freedom to make life choices
X_5 = Generosity
X_6 = Perceptions of corruption

Analyzing the original data



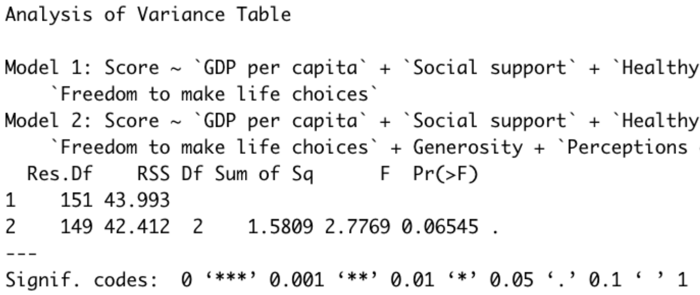
Graph 1. Scatter plot matrix of original data

According to Graph 1, the variables GDP per capita, social support, healthy life expectancy, and freedom to make life choices have somewhat positive linear associations with each other. However, the variables generosity and perceptions of corruption do not seem to have any linear associations with each other, or they have very weak positive linear associations.



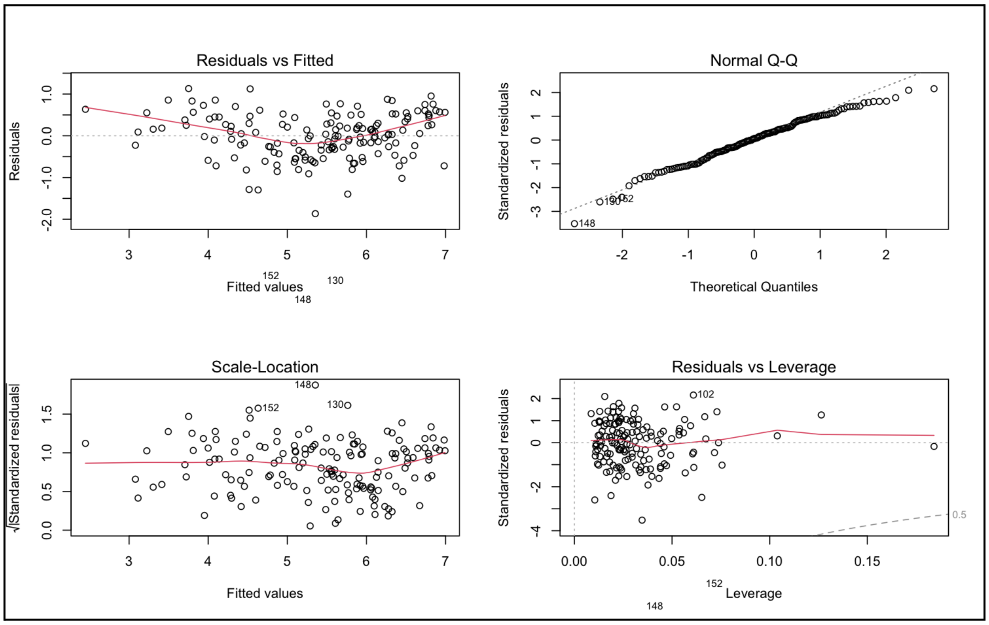
```
Call:
lm(formula = Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` +
    `Freedom to make life choices` + Generosity + `Perceptions of corruption`)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.75304 -0.35306  0.05703  0.36695  1.19059 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    1.7952     0.2111   8.505 1.77e-14 ***
`GDP per capita`               0.7754     0.2182   3.553 0.000510 ***
`Social support`               1.1242     0.2369   4.745 4.83e-06 ***
`Healthy life expectancy`      1.0781     0.3345   3.223 0.001560 ** 
`Freedom to make life choices` 1.4548     0.3753   3.876 0.000159 ***
Generosity                     0.4898     0.4977   0.984 0.326709    
`Perceptions of corruption`    0.9723     0.5424   1.793 0.075053 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5335 on 149 degrees of freedom
Multiple R-squared:  0.7792,	Adjusted R-squared:  0.7703 
F-statistic: 87.62 on 6 and 149 DF,  p-value: < 2.2e-16
```

According to the summary of the data, the overall linear model is significant, since the p-value of ANOVA is < 2.2e-16, which is less than 0.05. However, the slopes of the variables “Generosity” and “Perceptions of corruption” are not significant, with p-values greater than 0.05. Also, according to Appendix A, the added-variable plots of the variables “Generosity” and “Perceptions of corruption” have relatively flat lines, meaning they lack statistical significance to the model. Thus, we conduct a partial F test to determine whether or not to retain these two variables in the model.



```
Analysis of Variance Table

Model 1: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` + 
    `Freedom to make life choices`
Model 2: Score ~ `GDP per capita` + `Social support` + `Healthy life expectancy` + 
    `Freedom to make life choices` + Generosity + `Perceptions of corruption`
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1    151 43.993                              
2    149 42.412  2    1.5809 2.7769 0.06545 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

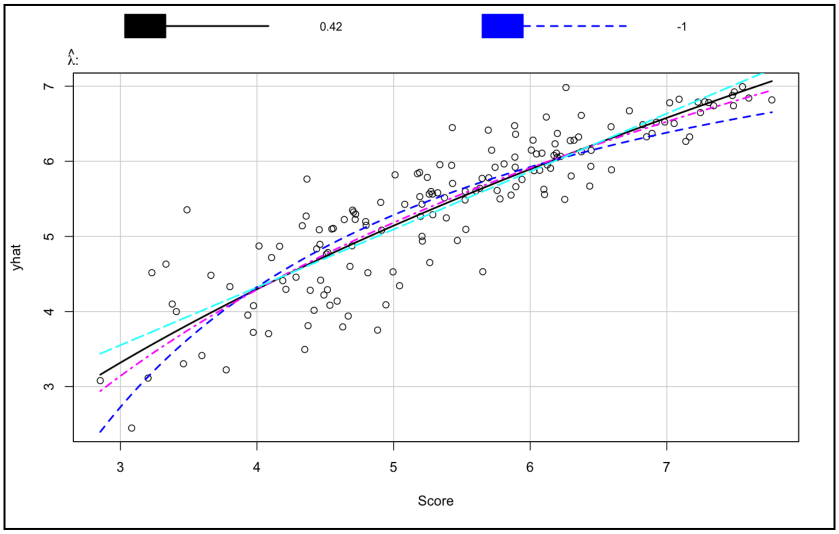
The result from the partial F test has a p-value of 0.06545, which is greater than 0.05. Thus, we fail to reject the null hypothesis, and there is no sufficient evidence against the reduced model in favor of the full model. As a result, we use the reduced model: Y ~ $X_1 + X_2 + X_3 + X_4$.

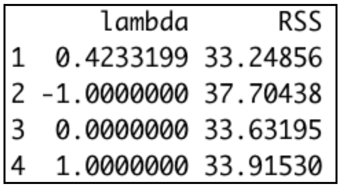


Graph 2. Diagnostic plots of the reduced model

According to Graph 2, the linearity assumption of the regression is violated, as the red line in the Residuals vs Fitted plot has a curved shape and is not equal to zero. In addition, certain points on the Normal Q-Q plot do not follow the trend of the diagonal straight line, violating the assumption of the normality of the error term. Thus, we execute a few transformations to find potential improvements in the model assumptions.

Transformation approach 1: Use inverse response plot to find the transformation for the response variable



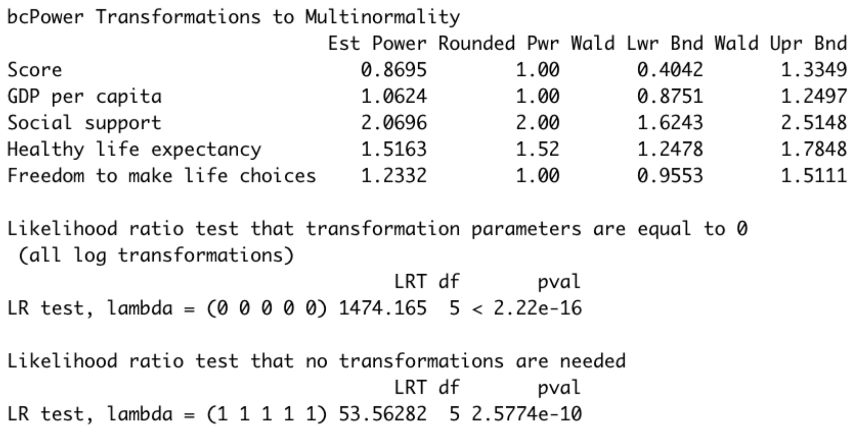


```
      lambda      RSS
1  0.4233199 33.24856
2 -1.0000000 37.70438
3  0.0000000 33.63195
4  1.0000000 33.91530
```

Graph 3. Inverse response plot of reduced model

According to Graph 3, the best choice of λ is 0.42. Thus, we transform the model into Y^0.42 ~ X_1 + X_2 + X_3+ X_4. According to Appendix B, the overall model is valid, with an ANOVA p-value of < 2.2e-16. All of the predictors are also significant, with p-values less than 0.05. After the transformation, according to Appendix C, there is a slight improvement in the linearity assumption of the model, as the red line in the Residuals vs Fitted plot becomes less curved. There is still space for improvement of the model assumptions, thus we see if there are better forms of transformation.

Transformation approach 2: Transform X and Y simultaneously using Box-cox method



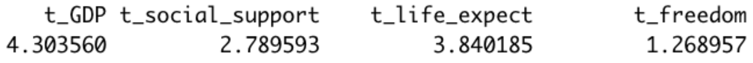
```
bcPower Transformations to Multinormality
                             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Score                           0.8695        1.00       0.4042       1.3349
GDP per capita                  1.0624        1.00       0.8751       1.2497
Social support                  2.0696        2.00       1.6243       2.5148
Healthy life expectancy         1.5163        1.52       1.2478       1.7848
Freedom to make life choices    1.2332        1.00       0.9553       1.5111

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                                  LRT df       pval
LR test, lambda = (0 0 0 0 0) 1474.165  5 < 2.22e-16

Likelihood ratio test that no transformations are needed
                                  LRT df       pval
LR test, lambda = (1 1 1 1 1) 53.56282  5 2.5774e-10
```

According to the summary of the Box-cox method, we transform the model using the rounded powers: Y ~ X_1 + X_2^2 + X_3^1.52 + X_4. According to Appendix D, the overall model is valid, with an ANOVA p-value of < 2.2e-16. All of the predictors are also significant, with p-values less than 0.05. After the transformation, according to Appendix E, the assumptions linearity has slightly improved compared to the untransformed model, with the red line in the Residual vs Fitted values graph being less curved. Also, the assumptions of constant variance of the error term are also improved, with the data points in the √|Standardized residuals| vs Fitted values graph more randomly and evenly scattered. In addition, the likelihood ratio tests, with p-values less than 0.05, do not suggest using a log transformation or no transformation of the model.

Comparing the untransformed reduced model (Y ~ X_1 + X_2 + X_3 + X_4), transformed model using the inverse response plot (Y^0.42 ~ X_1 + X_2 + X_3+ X_4), and transformed model using the box-cox method (Y ~ X_1 + X_2^2 + X_3^1.52 + X_4), we see that the transformed model using the box-cox method is the most valid in terms of satisfying the assumptions of the model (linearity, zero mean, normality, and constant variance of the error terms). Thus, we choose to use the model Y ~ X_1 + X_2^2 + X_3^1.52 + X_4.

Variable Selection

After choosing Y ~ X_1 + X_2^2 + X_3^1.52 + X_4 (the box-cox method transformed model) as our model, we would like to see if there is any multicollinearity between the predictor variables that causes their slopes to be poorly estimated.



```
   t_GDP t_social_support    t_life_expect        t_freedom 
4.303560         2.789593         3.840185         1.268957 
```

Referring to the VIF values above, none of the predictors have a VIF greater than 5. So there is not a multicollinearity issue in the model, and variable selection is not necessary. However, we would still like to verify this statement by using the following approaches:

Approach 1: All possible subsets

In *Appendix F*, we explore the optimal model for all sizes. To choose the best among the models, we compare their goodness of fit criteria as seen in *Appendix G*. In the adjusted-R^2 vs subset size plot, we can see that the model with four predictors has the highest R^2_{adj} value. Similarly, in the subsequent plots of BIC, AIC, and AICc, we observe that the model with four predictors seems to have the smallest values. Thus, from this approach, we conclude that the best model is Y ~ X_1 + X_2 + X_3 + X_4.

Approach 2: Stepwise Regression

Using a backward stepwise approach, we observe the optimal model by using the AIC and BIC method in *Appendix H and I*. The R output stops the approach at its first step, indicating that using four predictors is the most optimal for the data.

Using a forward stepwise approach, we observe the optimal model by using the AIC and BIC method in *Appendix J and K*. Here, the regression starts with the null model (intercept only). We see that the R output stops the approach at its last step, indicating that using four variables is the most optimal for the data.

Discussion and Results

R results

Based on the results of our hypothesis testing and variable selection methods, we conclude that the optimal model is

> Score ~ GDP per capita + Social Support + Healthy Life Expectancy + Freedom to make life choices.

with the box-cox method transformation.

This model has been validated via diagnostic plots and numerical results. From this, we understand that the happiness score of each country is mostly related to the financial and social well-being of the citizens, especially social support.

GDP per capita: Economic prosperity is essential for ensuring a comfortable life. It can influence happiness by providing individuals with access to resources and opportunities.

Social Support: Social support encompasses a reliable social network and trust in the public system. Access to affordable public transportation, healthcare, and education are examples of government-aided social support. People with robust social networks tend to experience reduced stress levels and improved mental health, which is crucial to their happiness level.

Healthy Life Expectancy: Countries with better healthcare systems and medical services allow their citizens to enjoy a higher quality of life. Longer and healthier lives enable individuals to engage in meaningful pursuits, contributing to their overall happiness.

Freedom to Make Life Choices: Societies that prioritize individual freedoms and human rights create an environment where people can pursue their aspirations and shape their lives according to their own desires.

According to the 2019 World Happiness Report, half of the countries that hold the top ten highest happiness scores are located in Northern Europe. This is not surprising because the countries in the Nordic region are known for adopting a societal framework known as the “Nordic model.” The Nordic model incorporates elements of both capitalism and socialism, meaning that the citizens enjoy a free market economy and social benefits at the same time (“Nordic Model: Comparing The Economic System to the U.S.”). The high living standards and low-income disparities in these countries are reflected in the happiness score of each nation. Comparatively, the countries with the lowest scores are mostly located in areas plagued by war, terrorism, and low living standards. With troubling domestic affairs, the governments in those countries are unable to provide social support for their citizens.

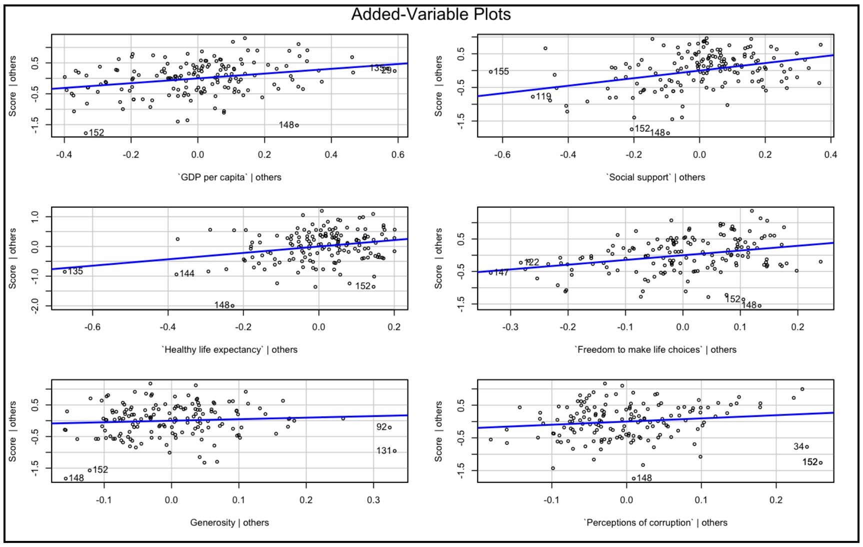
Limitations

The final model has the best diagnostic plots compared to the other models. However, it is impossible to entirely satisfy the model assumption, since this analysis uses real-life data, which is normal to have errors. Also, after transforming the model using the box-cox method, it is hard to interpret the slope coefficients of the predictor. While the original direct relationship between the response variable and predictor variables is not preserved, the transformed data’s coefficients still provide the magnitude and direction of the relationship of the variables. We can also see which predictors have more influence on the response variable based on the p-values, which indicates the predictors’ significance. In addition, the R^2 and adjusted R^2 of the models were constantly in the 0.7 to 0.8 range, which is not considered particularly high (being greater than 0.9). This may be because of the error of the real-life data, or there might be more suitable models for the data. However, this is not a major concern of the model, since the final model is significant according to the F-test.

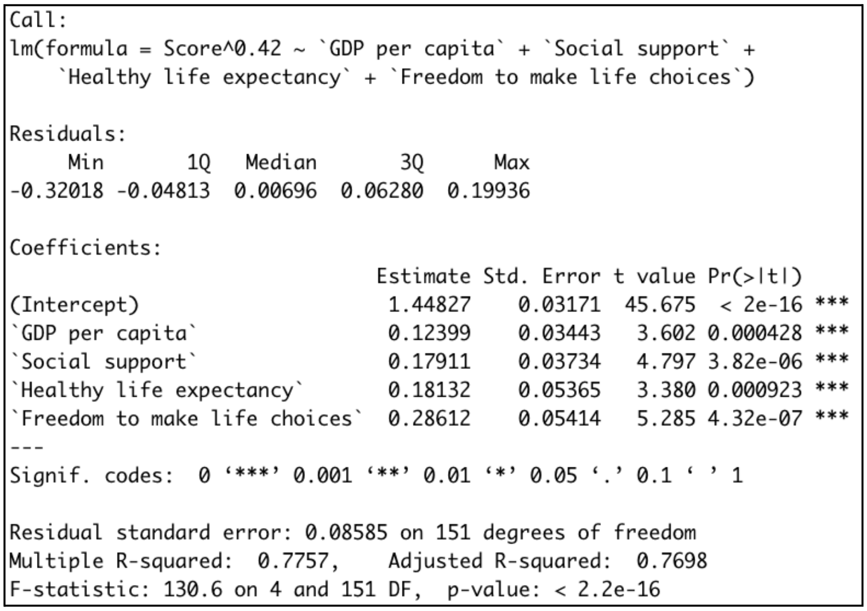
To overcome the difficulty of interpreting the slope coefficients of the box-cox method transformed data, in the future, we can use a log transformation if it improves the model and its assumptions, since a log transformation allows us to explain the relationship between the predictor and response variables in terms of percentage changes. We can also try approaches other than linear regression to see if the R^2 and adjusted R^2 will improve.

Appendix

Appendix A: *Added-variable plots of $Y \sim X_1 + X_2 + X_3 + X_4 + X_5 + X_6$ (original data)*

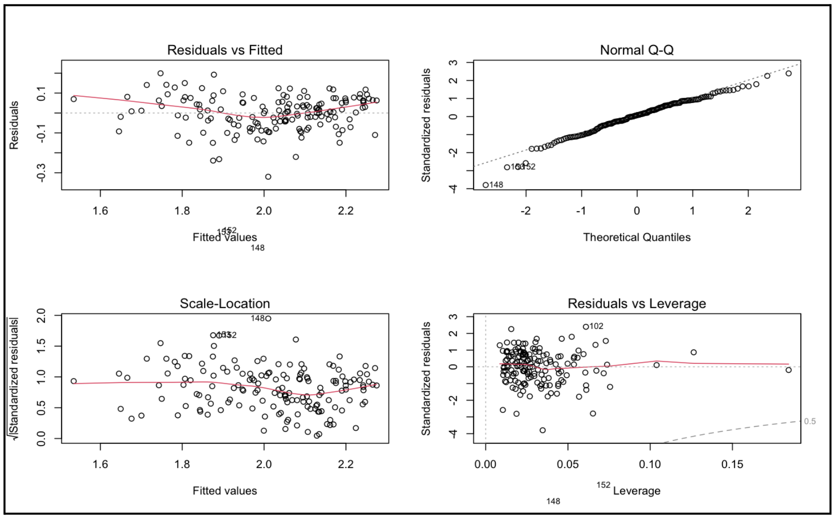


Appendix B: *Summary of Score^0.42 ~ GDP per capita + Social support + Healthy life expectancy + Freedom to make life choices.*

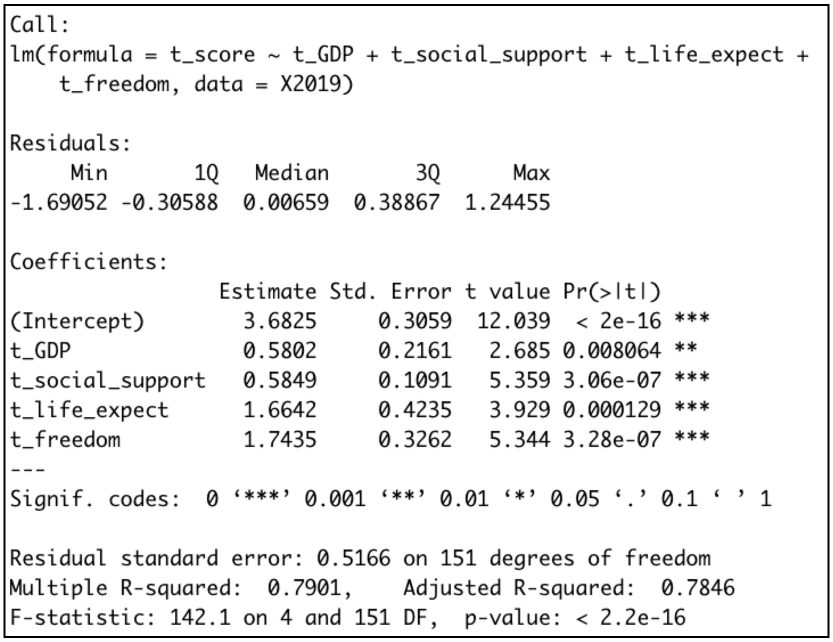


```
Call:
lm(formula = Score^0.42 ~ `GDP per capita` + `Social support` +
    `Healthy life expectancy` + `Freedom to make life choices`)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32018 -0.04813  0.00696  0.06280  0.19936 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.44827    0.03171  45.675  < 2e-16 ***
`GDP per capita`                0.12399    0.03443   3.602 0.000428 ***
`Social support`                0.17911    0.03734   4.797 3.82e-06 ***
`Healthy life expectancy`       0.18132    0.05365   3.380 0.000923 ***
`Freedom to make life choices`  0.28612    0.05414   5.285 4.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08585 on 151 degrees of freedom
Multiple R-squared:  0.7757,	Adjusted R-squared:  0.7698 
F-statistic: 130.6 on 4 and 151 DF,  p-value: < 2.2e-16
```

Appendix C: *Diagnostic plots of Score^0.42 ~ GDP per capita + Social support + Healthy life expectancy + Freedom to make life choices.*

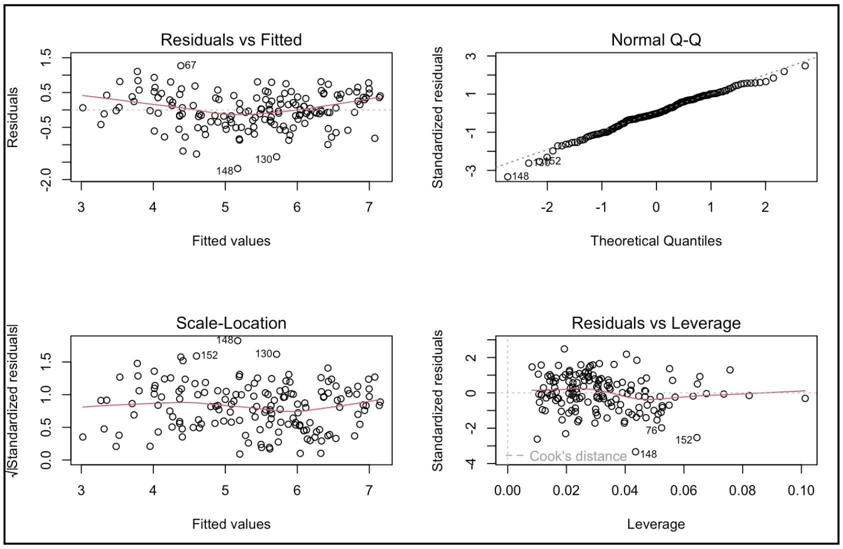


Appendix D: *Summary of Score ~ GDP per capita + Social support^2 + Healthy life expectancy^1.52 + Freedom to make life choices.*

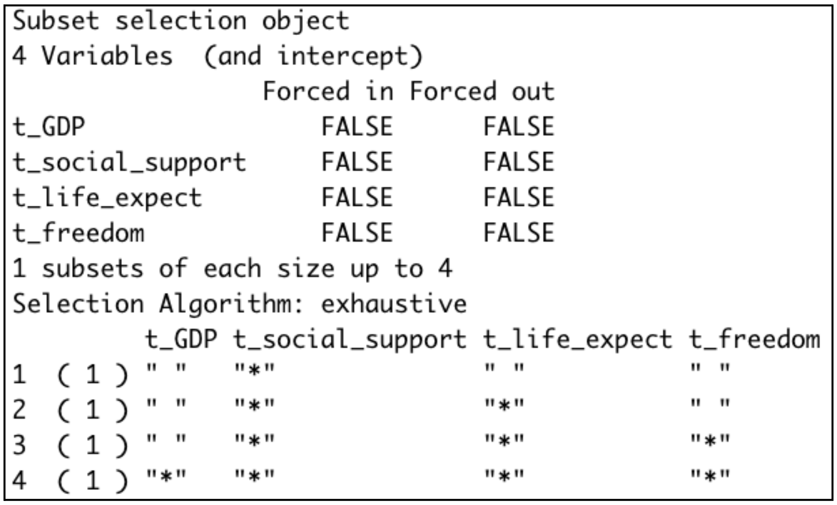


```
Call:
lm(formula = t_score ~ t_GDP + t_social_support + t_life_expect +
    t_freedom, data = X2019)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.69052 -0.30588  0.00659  0.38867  1.24455 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3.6825     0.3059  12.039  < 2e-16 ***
t_GDP              0.5802     0.2161   2.685 0.008064 ** 
t_social_support   0.5849     0.1091   5.359 3.06e-07 ***
t_life_expect      1.6642     0.4235   3.929 0.000129 ***
t_freedom          1.7435     0.3262   5.344 3.28e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5166 on 151 degrees of freedom
Multiple R-squared:  0.7901,	Adjusted R-squared:  0.7846 
F-statistic: 142.1 on 4 and 151 DF,  p-value: < 2.2e-16
```

Appendix E: *Diagnostic plot of Score ~ GDP per capita + Social support^2 + Healthy life expectancy^1.52 + Freedom to make life choices.*

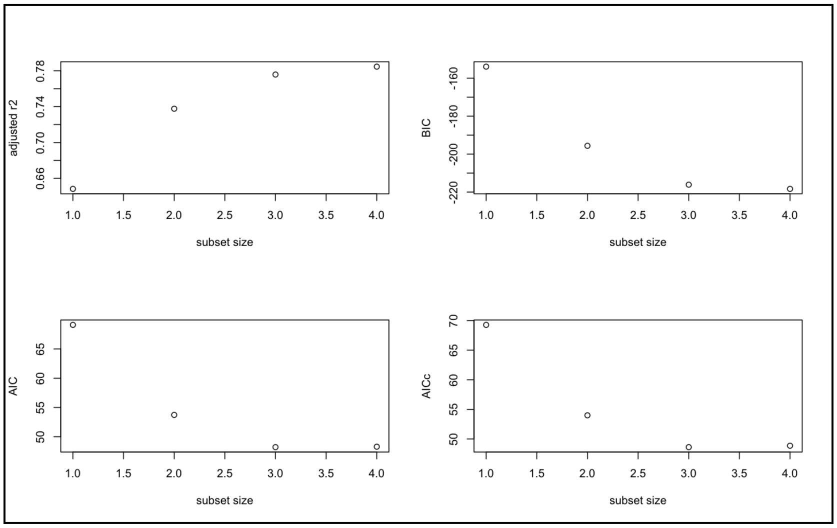


Appendix F: *All possible subsets of Score ~ GDP per capita + Social support^2 + Healthy life expectancy^1.52 + Freedom to make life choices.*

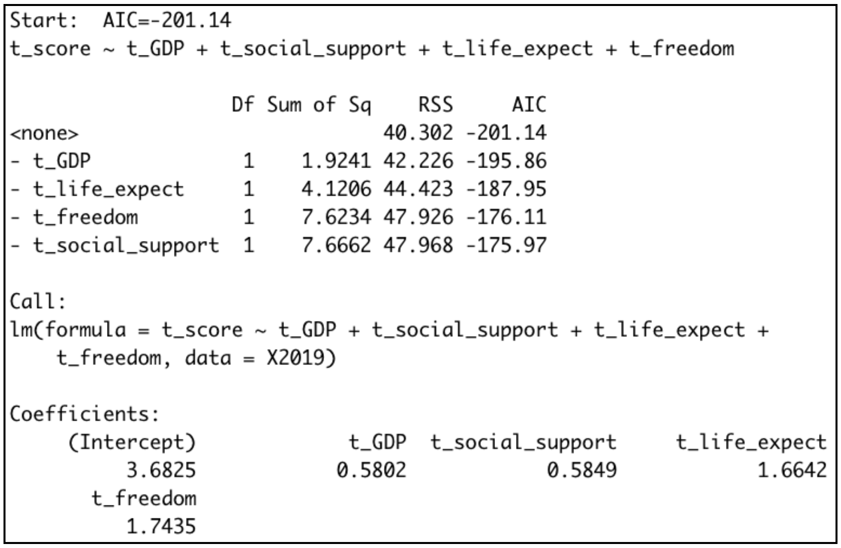


```
Subset selection object
4 Variables  (and intercept)
                 Forced in Forced out
t_GDP                FALSE      FALSE
t_social_support     FALSE      FALSE
t_life_expect        FALSE      FALSE
t_freedom            FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         t_GDP t_social_support t_life_expect t_freedom
1  ( 1 ) " "   "*"              " "           " "
2  ( 1 ) " "   "*"              "*"           " "
3  ( 1 ) " "   "*"              "*"           "*"
4  ( 1 ) "*"   "*"              "*"           "*"
```

Appendix G: *Subset sizes vs adjusted R^2, AIC, AICc, BIC*

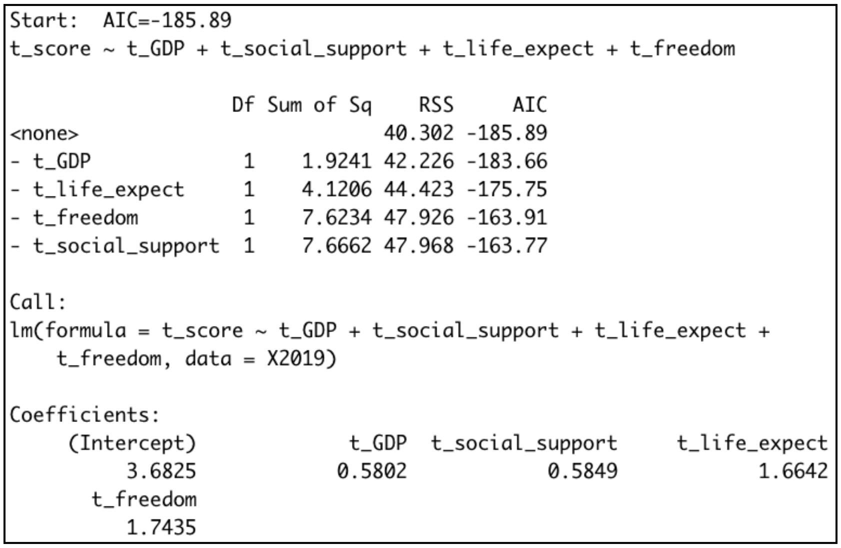


Appendix H: *Backward AIC method*



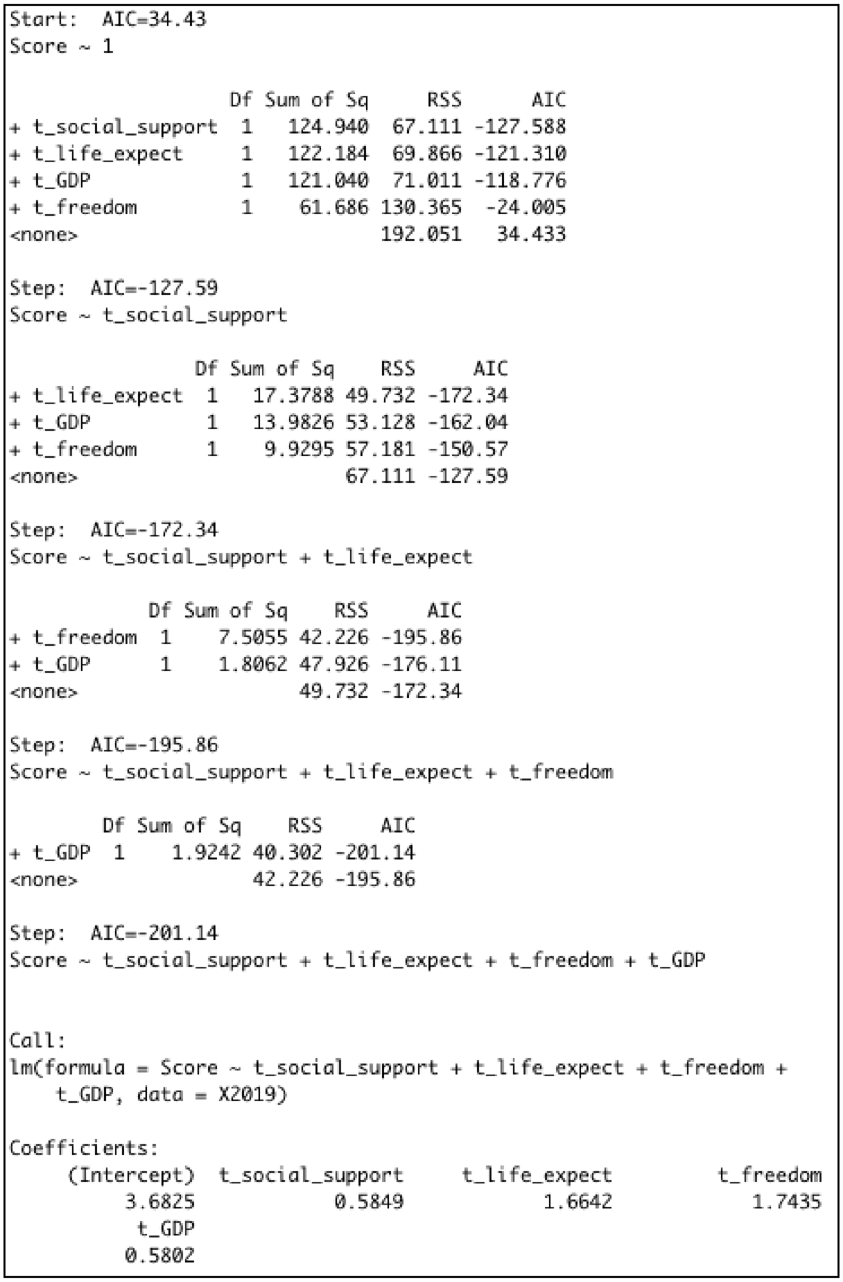
```
Start:  AIC=-201.14
t_score ~ t_GDP + t_social_support + t_life_expect + t_freedom

                   Df Sum of Sq    RSS     AIC
<none>                          40.302 -201.14
- t_GDP             1    1.9241 42.226 -195.86
- t_life_expect     1    4.1206 44.423 -187.95
- t_freedom         1    7.6234 47.926 -176.11
- t_social_support  1    7.6662 47.968 -175.97

Call:
lm(formula = t_score ~ t_GDP + t_social_support + t_life_expect +
    t_freedom, data = X2019)

Coefficients:
     (Intercept)             t_GDP  t_social_support     t_life_expect
          3.6825            0.5802            0.5849            1.6642
       t_freedom
          1.7435
```

Appendix I: *Backward BIC method*



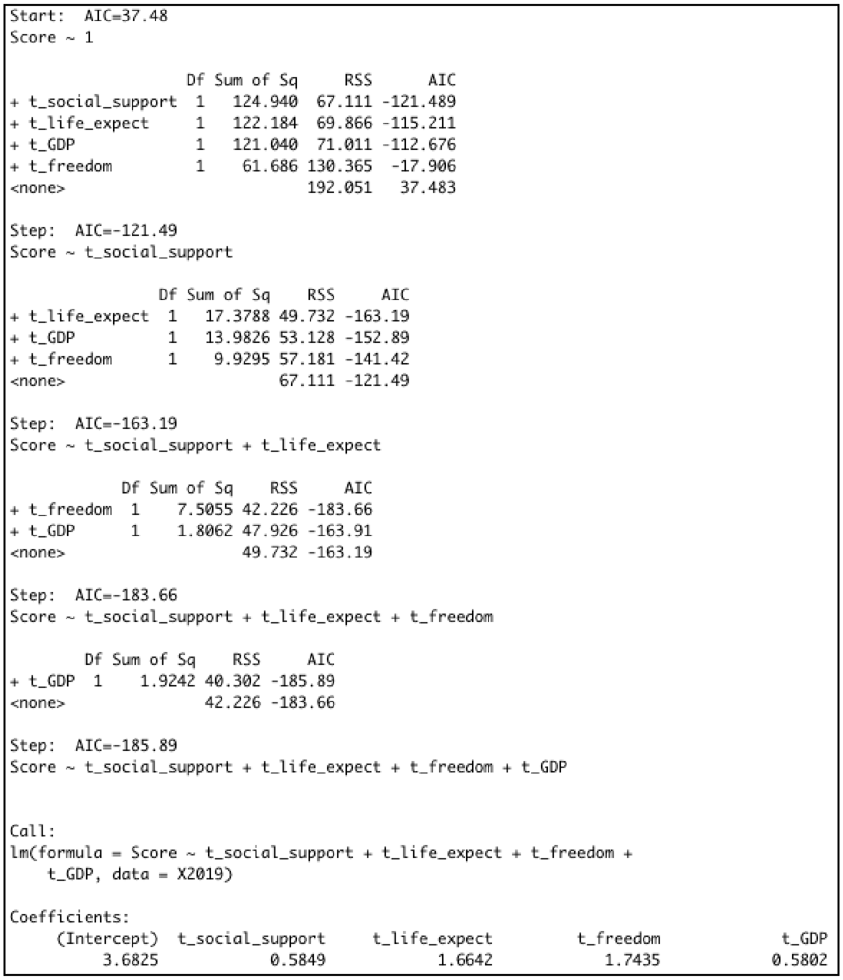
```
Start:  AIC=-185.89
t_score ~ t_GDP + t_social_support + t_life_expect + t_freedom

                   Df Sum of Sq    RSS     AIC
<none>                          40.302 -185.89
- t_GDP             1    1.9241 42.226 -183.66
- t_life_expect     1    4.1206 44.423 -175.75
- t_freedom         1    7.6234 47.926 -163.91
- t_social_support  1    7.6662 47.968 -163.77

Call:
lm(formula = t_score ~ t_GDP + t_social_support + t_life_expect +
    t_freedom, data = X2019)

Coefficients:
     (Intercept)             t_GDP  t_social_support     t_life_expect
          3.6825            0.5802            0.5849            1.6642
       t_freedom
          1.7435
```

Appendix J: *Forward AIC method*



```
Start:  AIC=34.43
Score ~ 1

                   Df Sum of Sq     RSS      AIC
+ t_social_support  1   124.940  67.111 -127.588
+ t_life_expect     1   122.184  69.866 -121.310
+ t_GDP             1   121.040  71.011 -118.776
+ t_freedom         1    61.686 130.365  -24.005
<none>                          192.051   34.433

Step:  AIC=-127.59
Score ~ t_social_support

                Df Sum of Sq    RSS     AIC
+ t_life_expect  1   17.3788 49.732 -172.34
+ t_GDP          1   13.9826 53.128 -162.04
+ t_freedom      1    9.9295 57.181 -150.57
<none>                       67.111 -127.59

Step:  AIC=-172.34
Score ~ t_social_support + t_life_expect

            Df Sum of Sq    RSS     AIC
+ t_freedom  1    7.5055 42.226 -195.86
+ t_GDP      1    1.8062 47.926 -176.11
<none>                   49.732 -172.34

Step:  AIC=-195.86
Score ~ t_social_support + t_life_expect + t_freedom

        Df Sum of Sq    RSS     AIC
+ t_GDP  1    1.9242 40.302 -201.14
<none>               42.226 -195.86

Step:  AIC=-201.14
Score ~ t_social_support + t_life_expect + t_freedom + t_GDP


Call:
lm(formula = Score ~ t_social_support + t_life_expect + t_freedom + 
    t_GDP, data = X2019)

Coefficients:
     (Intercept)  t_social_support     t_life_expect         t_freedom  
          3.6825            0.5849            1.6642            1.7435  
           t_GDP  
          0.5802  
```

Appendix K: *Forward BIC method*



```
Start:  AIC=37.48
Score ~ 1

                   Df Sum of Sq     RSS      AIC
+ t_social_support  1   124.940  67.111 -121.489
+ t_life_expect     1   122.184  69.866 -115.211
+ t_GDP             1   121.040  71.011 -112.676
+ t_freedom         1    61.686 130.365  -17.906
<none>                          192.051   37.483

Step:  AIC=-121.49
Score ~ t_social_support

                Df Sum of Sq    RSS     AIC
+ t_life_expect  1   17.3788 49.732 -163.19
+ t_GDP          1   13.9826 53.128 -152.89
+ t_freedom      1    9.9295 57.181 -141.42
<none>                       67.111 -121.49

Step:  AIC=-163.19
Score ~ t_social_support + t_life_expect

            Df Sum of Sq    RSS     AIC
+ t_freedom  1    7.5055 42.226 -183.66
+ t_GDP      1    1.8062 47.926 -163.91
<none>                   49.732 -163.19

Step:  AIC=-183.66
Score ~ t_social_support + t_life_expect + t_freedom

        Df Sum of Sq    RSS     AIC
+ t_GDP  1    1.9242 40.302 -185.89
<none>               42.226 -183.66

Step:  AIC=-185.89
Score ~ t_social_support + t_life_expect + t_freedom + t_GDP


Call:
lm(formula = Score ~ t_social_support + t_life_expect + t_freedom + 
    t_GDP, data = X2019)

Coefficients:
     (Intercept)  t_social_support     t_life_expect         t_freedom             t_GDP  
          3.6825            0.5849            1.6642            1.7435            0.5802  
```

References

Network, Sustainable Development Solutions. “World Happiness Report.” Kaggle, November 27, 2019. https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2019.csv.

“Nordic Model: Comparing The Economic System to the U.S.” *Investopedia*, https://www.investopedia.com/terms/n/nordic-model.asp. Accessed 6 August 2023.