Using Linear Regression to Study World Happiness Level

Hailey Lee | slee271@gmu.edu
Department of Computer Science | George Mason University

INTRODUCTION

Happiness is essential in individuals' lives. Happiness is a significant key to a better life in this world, including interactions with others, individuals' well-being, communities' development, and world productivity. There are many factors that affect individuals and the world's happiness, which also depends on where individuals reside.

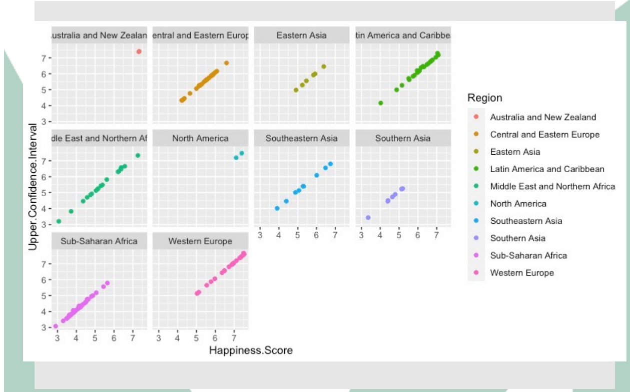


Figure 1 Relationship between Happiness Score and Upper Confidence Interval grouped by Region

RESEARCH QUESTION/GOAL

The goal of this study was to predict the happiness score of the countries that haven't been reported and update the happiness scores of the countries with updated status of the factors.

DATA AND DATA PREPARATION

The factors are Lower and Upper Confidence Interval, Economy, Family, Health, Freedom, Trust, and Generosity.

According to the data, published by the United Nations Sustainable Development Solutions Network, the top five countries with high happiness scores were Denmark, Switzerland, Iceland, Norway, and Norway. The bottom five were Burundi, Syria, Togo, Afghanistan, and Benin. The number of recorded countries is 158, and the range of the happiness score is approximately 4.6.

For the purpose of this study, I removed the variables Dystopia Residual and Happiness Rank.

DATA MODELING

I used the programming language R to analyze the correlation between happiness score and eight other factors. The correlation between the two variables is like this: 0.999 (Lower Confidence Interval), 0.999 (Upper Confidence Interval), 0.790 (Economy), 0.739 (Family), 0.765 (Health), 0.567 (Freedom), 0.462 (Trust), and 0.157 (Generosity). I repeated this analysis after I removed the last three lesser significant actors from these multiple regression models. To find out any patterns of this model, I made a residual graph and removed any lesser significant factors using square terms.

For Machine Learning Algorithms, I used a Supervised Learning Algorithm, which consists of independent and dependent variables and produces a linear equation with given input and desired output. Specifically, I will use Simple Linear Regression to show the relationship between two variables.

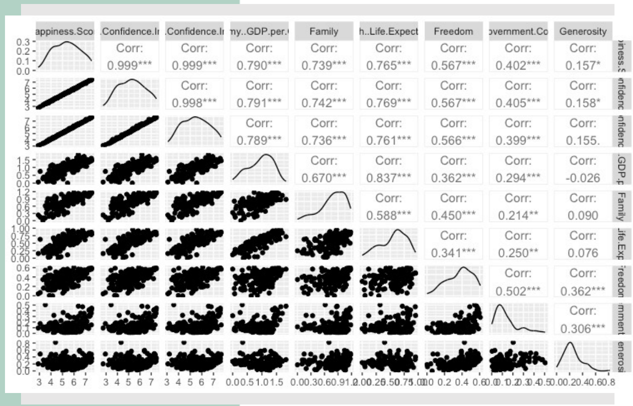
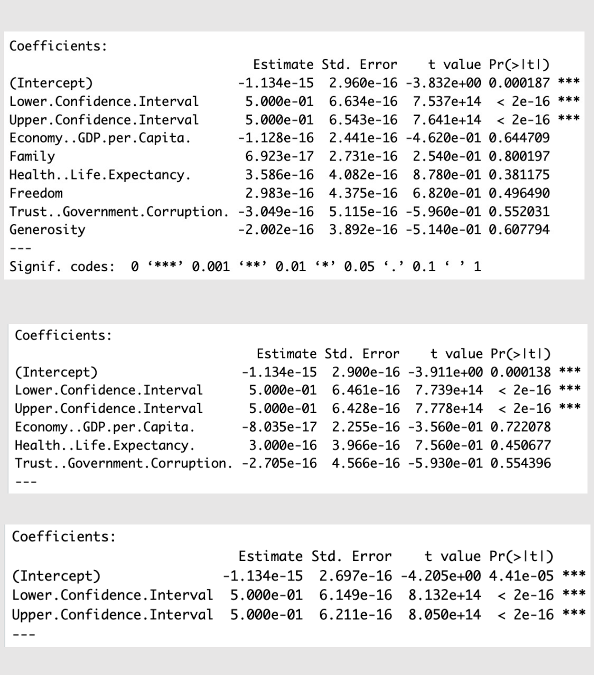


Figure 2 Correlation between each variables in data

DATA EVALUATION

For testing, I plan to use the Holdout Method, using two-thirds of the data for training and the last third of the data for testing.

For debugging, I will use the traceback() function to print out previously called functions before the error occurred and the recover() function to stop executing at the point when the error occurred. Also, to maximize efficiency, I plan to use the message function to print out the status of the modeling/debugging



```
Coefficients:
                              Estimate Std. Error    t value Pr(>|t|)    
(Intercept)                  -1.134e-15  2.960e-16 -3.832e+00 0.000187 ***
Lower.Confidence.Interval     5.000e-01  6.634e-16  7.537e+14  < 2e-16 ***
Upper.Confidence.Interval     5.000e-01  6.543e-16  7.641e+14  < 2e-16 ***
Economy..GDP.per.Capita.     -1.128e-16  2.441e-16 -4.620e-01 0.644709    
Family                        6.923e-17  2.731e-16  2.540e-01 0.800197    
Health..Life.Expectancy.      3.586e-16  4.082e-16  8.780e-01 0.381175    
Freedom                       2.983e-16  4.375e-16  6.820e-01 0.496490    
Trust..Government.Corruption. -3.049e-16  5.115e-16 -5.960e-01 0.552031    
Generosity                   -2.002e-16  3.892e-16 -5.140e-01 0.607794    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Coefficients:
                              Estimate Std. Error    t value Pr(>|t|)    
(Intercept)                  -1.134e-15  2.900e-16 -3.911e+00 0.000138 ***
Lower.Confidence.Interval     5.000e-01  6.461e-16  7.739e+14  < 2e-16 ***
Upper.Confidence.Interval     5.000e-01  6.428e-16  7.778e+14  < 2e-16 ***
Economy..GDP.per.Capita.     -8.035e-17  2.255e-16 -3.560e-01 0.722078    
Health..Life.Expectancy.      3.000e-16  3.966e-16  7.560e-01 0.450677    
Trust..Government.Corruption. -2.705e-16  4.566e-16 -5.930e-01 0.554396    
---
```

```
Coefficients:
                           Estimate Std. Error    t value Pr(>|t|)    
(Intercept)              -1.134e-15  2.697e-16 -4.205e+00 4.41e-05 ***
Lower.Confidence.Interval  5.000e-01  6.149e-16  8.132e+14  < 2e-16 ***
Upper.Confidence.Interval  5.000e-01  6.211e-16  8.050e+14  < 2e-16 ***
---
```

Figure 3 R code of process of finding the most accurate regression model

CONCLUSIONS/DISCUSSION

According to the result, the decision on happiness level using the Upper and Lower Confidence Interval was the best approach. However, it also depended on the region; therefore, the future study will be divided into regions and build a model for each.

REFERENCES:

Ray, S. (2023, May 26). *Commonly used machine learning algorithms | Data science*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Speegle, D., & Clair, B. (2021). *Probability, statistics, and data: A fresh approach using R*. CRC Press.

Torgo, L. (2016). *Data mining with R: Learning with case studies* (2nd ed.). CRC Press.